MODELLING RISKS OF HOSPITAL MORTALITY FOR CRITICALLY ILL PATIENTS

ROWENA WONG SYN YIN

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF ECONOMICS AND ADMINISTRATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: ROWENA WONG SYN YIN

Matric No: EHA080021

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

MODELLING RISKS OF HOSPITAL MORTALITY FOR CRITICALLY ILL PATIENTS

Field of Study: APPLIED STATISTICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Intensive care unit (ICU) prognostic models can be used to predict mortality outcomes for critically ill patients who require intensive treatment due to the severity of their illness. These physiological and statistical-based models stratify patients according to their severity of illness and provide an objective approach in predicting hospital mortality risks. These models are useful tools in assisting clinicians in decision making, interpretation of diagnosis and prescription of appropriate treatment options to patients. They can also be effectively used for benchmarking purposes to evaluate and compare the clinical performances of different ICUs and assist hospital administration in making informed changes in resource allocations. Although these models are predominantly used in developed countries, they are not that popular in developing countries due to costs, facilities and resources considerations. In this study, the advantages, limitations and evolutions of three selected well-established ICU prognostic systems were reviewed and discussed. The Acute Physiology and Chronic Health Evaluation (APACHE IV) model was chosen as the reference model in this study due to its promising potential as a suitable benchmarking tool. The first objective of this study is to investigate the validity of APACHE IV model in predicting mortality risk in a Malaysian ICU. A prospective independent observational study was conducted at a single-centre multidisciplinary ICU in Hospital Sultanah Aminah Johor Bahru (HSA ICU). External validation of APACHE IV involved a cohort of 916 admissions to HSA ICU in the year 2009. APACHE IV was found to be not suitable for application in HSA ICU. Although the model exhibited good discrimination, calibration was observed to be poor. The model overestimated risk of death in HSA ICU, especially for mid- to high- risk patient groups. The model's lack of fit was mainly attributed to differences in case mix and patient management between APACHE IV and HSA ICU. The second objective of this

research involves investigation of the significant factors that affect mortality risk in HSA ICU and development of a prognostic model that is suitable for application in HSA ICU. Bayesian Markov Chain Monte Carlo and decision tree approaches were explored as alternative methods in the modelling of ICU risk of death, where five different types of Bayesian models and a decision tree model were proposed in this research. Although the performance of the decision tree model was comparable to the Bayesian models, it was not as informative as the Bayesian models, especially in predicting individual patient mortality risk. One of the Bayesian models was chosen as the best model to be used as the future reference model in HSA ICU. This model comprises seven variables (age, gender, Acute Physiological Score (APS), absence of Glasgow Coma Scale score, mechanical ventilation, presence of chronic health and ICU admission diagnoses) that are readily available in any intensive care unit setting. This research has shown the promising potential of the Bayesian approach as an alternative in the analysis and modelling of ICU mortality risks.

ABSTRAK

Model prognostik boleh digunakan untuk meramalkan risiko kematian untuk pesakit kritikal yang memerlukan rawatan intensif di unit rawatan rapi. Pembinaan model prognostik adalah berdasarkan komponen fisiologi dan aplikasi statistik. Model prognostik boleh digunakan untuk penstrataan pesakit mengikut tahap kritikal penyakit mereka. Di samping itu, model prognostik menawarkan satu pendekatan objektif dalam ramalan risiko kematian di dalam hospital. Model-model ini boleh membantu doktor dalam membuat keputusan, tafsiran diagnosis dan preskripsi tentang pilihan rawatan yang paling sesuai untuk pesakit. Mereka juga boleh digunakan sebagai penanda aras untuk menilai dan membandingkan pencapaian klinikal di unit rawatan rapi, serta membantu pentadbiran hospital dalam membuat keputusan tentang peruntukan sumber. Walaupun model-model ini kebanyakannya digunakan di negara-negara maju seperti Amerika Syarikat, Eropah dan Australia, kekangan kos, kemudahan dan sumber menyebabkan model-model ini tidak begitu popular di negara-negara yang sedang membangun. Perbandingan tentang ciri-ciri, kelemahan dan evolusi tiga jenis sistem prognostik popular yang mantap telah dibincangkan di dalam kajian ini. Model Akut Fisiologi dan Penilaian Kesihatan Kronik (APACHE IV) dipercayai mempunyai potensi yang baik sebagai penanda aras dan telah dipilih sebagai model rujukan dalam kajian ini. Satu kajian pemerhatian bebas telah dijalankan di unit rawatan rapi di Hospital Sultanah Aminah Johor Bahru (ICU HSA). Objektif pertama kajian ini adalah untuk menyiasat kesahihan dan kesesuaian model APACHE IV dalam meramalkan risiko kematian di ICU HSA. Pengesahan model APACHE IV melibatkan 916 pesakit yang dimasukkan ke ICU HSA pada tahun 2009. APACHE IV telah didapati tidak sesuai untuk digunakan di ICU HSA. Walaupun model ini menunjukkan diskriminasi baik, kalibrasi model didapati tidak memuaskan. Model ini terlebih menganggar risiko

kematian dalam ICU HSA, terutama bagi kumpulan pesakit yang mempunyai risiko sederhana ke peringkat yang lebih tinggi. Keputusan ini disebabkan oleh perbezaan dalam campuran kes dan pengurusan pesakit antara APACHE IV dan HSA ICU. Objektif kedua kajian ini melibatkan penyiasatan faktor-faktor yang mempengaruhi risiko kematian di ICU HSA dan pembinaan model ramalan yang sesuai untuk ICU HSA. Kaedah rantaian Markov Monte Carlo Bayesan dan pokok keputusan telah digunakan sebagai pendekatan alternatif dalam pemodelan risiko kematian, di mana lima jenis model Bayesan dan satu model pokok keputusan telah dicadangkan. Walaupun prestasi model pokok keputusan adalah setanding dengan model Bayesan, model pokok keputusan kurang sesuai digunakan untuk ramalan risiko kematian bagi pesakit individu. Salah satu model Bayesan disyorkan sebagai model yang terbaik untuk dijadikan rujukan masa depan dalam ICU HSA berdasarkan prestasi secara keseluruhan. Model ini mengandungi tujuh pembolehubah (umur, jantina, skor akut fisiologi (APS), ketiadaan skor Skala Glasgow Coma, pengudaraan mekanikal, kesihatan kronik dan diagnosis kemasukan unit rawatan rapi) yang mudah diperolehi di mana-mana unit rawatan rapi. Kajian ini telah berjaya menunjukkan potensi pendekatan rantaian Markov Monte Carlo Bayesan sebagai alternatif dalam analisis dan pemodelan risiko kematian dalam unit rawatan rapi.

ACKNOWLEDGEMENTS

I would like express my deepest appreciation to my supervisor, Professor Dr. Noor Azina Ismail, for her valuable guidance, understanding, continuous encouragement and constructive suggestions throughout the whole journey in completing this study.

This journey would not have been possible without the support of my family. I am especially grateful to all of my family members for their moral and financial support, and in allowing me the freedom to pursue my dreams.

I would also like to thank the examiners for their insightful suggestions that helped to improve the overall quality of this thesis.

Finally, I would like to say a special heartfelt thank you to my good friend, Dr. Dharini A/P Pathmanathan, for her constant prayers, help and spiritual support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

APACHE	:	Acute Physiology and Chronic Health Evaluation
APS	:	Acute Physiology Score
AUC	:	area under receiver operating characteristic curve
BGR	:	Brooks-Gelman-Rubin
BUN	:	blood urea nitrogen
CABG	:	coronary artery bypass graft
CART	:	classification and regression tree
CDSS	:	clinical decision support system
CHAID	:	chi-squared automatic interaction detector
CODA	:	convergence diagnosis and output analysis
DIC	:	deviance information criterion
FiO_2	:	fraction of inspired oxygen
GCS	:	Glasgow Coma Scale
HL	:	Hosmer-Lemeshow
HSA	:	Hospital Sultanah Aminah
ICU	:	intensive care unit
LOWESS	:	locally weighted scatterplot smoothing
MC	:	Monte Carlo
MCMC	:	Markov Chain Monte Carlo
MLE	:	maximum likelihood estimation
MPM	:	Mortality Probability Models
PaO_2	:	partial pressure of oxygen in arterial blood
ROC	:	receiver operating characteristic
SAPS	:	Simplified Acute Physiology Score

SD	:	standard deviation
SE	:	standard error
SMR	:	Standardised Mortality Ratio
SOFA	:	Sequential Organ Failure Assessment
WBC	:	white blood cell

CHAPTER 1: INTRODUCTION

1.1 Severity of illness scoring systems and intensive care unit prognostic models

Clinical decision rules are important to aid physicians in determining patients' diagnosis and prognosis. These rules are useful in situations where decision making is complex and when the clinical stakes are high (McGinn et al., 2000). Nowadays, with the advent of technology and wide access to computer systems, clinical decision rules are usually incorporated in clinical decision support systems. A clinical decision support system (CDSS) is defined as any electronic or non-electronic system that is designed to aid clinical decision making, whereby the characteristics of patients are matched to a computerised knowledge base and used to generate patient-specific assessments (Hunt, Haynes, Hanna, & Smith, 1998). Application of CDSS is not restricted to specific areas of medical care and most of these systems are designed for use in a heterogeneous environment. These systems have been widely used to improve drug prescriptions, provide computerised reminders for preventive care and assist in disease management such as hypertension, diabetes or acquired immunodeficiency syndrome (AIDS) (Hunt et al., 1998). CDSS is also applied in paediatric critical care and has been proven to reduce the rates of wrong drug prescriptions, improve therapeutic dosage targets and reduce cost (Mullett, Evans, Christenson, & Dean, 2001).

Clinical decision support systems are also applied in the management of adult critical care to enhance patient care, improve patient outcomes and reduce errors (Purcell, 2005). In most hospitals, individual patient prognosis is commonly evaluated through the physician’s experience and clinical judgement. However, this approach has been criticised as being too subjective, judgemental and prone to bias. The reliability of this approach is also questionable since predictions drawn in such a subjective manner may not be consistent and reproducible over time (Cowen & Kelley, 1994). In recent

years, critical care intensivists are moving towards the use of severity of illness scoring systems and prognostic models as clinical decision support instruments. These systems are designed to improve the evaluation of patients' prognoses through a standard approach, and are competent in generating predicted outcomes that are objective, consistent and reproducible over time. Although the use of prognostic models is normally intended for prediction of group mortality, it can be extended for individual prognosis, provided factors such as impact of complications and response to therapy are taken into consideration (Zimmerman & Kramer, 2008). As such, they are useful in assisting clinicians to interpret diagnosis accurately and to prescribe appropriate treatment options. Other than being applied to assist in clinical decision-making, these models also serve as benchmarking tools to measure and compare the quality and performances of several ICUs for a given duration, as well as, within an individual unit over time. Hospital administrators can also benefit from application of these prognostic models because they can provide guidance in terms of resource allocation, such as whether there is a need to add more beds, or to adjust the staff-to-patient ratio in an ICU (Schwartz & Cullen, 1981).

The concept of severity of illness scoring systems first emerged in the 1980s, with the introduction of the Acute Physiology and Chronic Health Evaluation (APACHE) (Knaus, Zimmerman, Wagner, Draper, & Lawrence, 1981) and Simplified Acute Physiology Score (SAPS) (Le Gall et al., 1984) systems. In principle, these systems rely on the theory that data that are collected from critically ill patients can be used to predict their degree of severity of illness and the corresponding risk of death. As such, the systems take into account information such as patient characteristics and clinical variables such as age, physiological abnormalities, acute diagnoses and comorbidities.

Both systems adopt a data reduction technique that involves the use of a scoring approach to measure severity of illness through a patient's physiological abnormalities. Points (scores) are assigned to physiological variables that have been identified as important predictors of mortality risk, where higher points are given for abnormal physiological values. This scoring approach is based on the belief that increasingly severe physiological derangement of critically ill patients is associated with increasing mortality risk. APACHE and SAPS also take into consideration other variables that could potentially affect a patient's mortality risk, such as patient's age and presence of underlying chronic diseases. Age is often an important component of most severity of illness scoring systems because increasing chronological age has been found to be a significant factor in increasing the risk of hospital death after intensive care (Wagner, Knaus, & Draper, 1983). Thus, older patients are assigned higher points to reflect their higher risk of mortality. Similarly, patients with underlying chronic illnesses are also associated with a higher mortality risk, and are given higher points compared to those without underlying comorbidities.

Although APACHE and SAPS share a common approach in evaluating severity of illness in critically ill patients, they differ in certain aspects such as in the selection and weighting of variables. SAPS is considered an abbreviated version of APACHE, with fewer variables. Both systems require assessment of physiological variables within the first day after ICU admission, where points for the worst physiological indicators within this period are taken into consideration. A patient's degree of severity of illness is measured through an aggregate score that is calculated by combining the total points for age, physiological and chronic health components. The aggregate scores in APACHE and SAPS are used as reference in stratifying critically ill patients into different risk categories according to their severity of illness. Although these systems are useful for patient stratification purposes, they do not offer predictions of in-hospital mortality risk.

Lemeshow, Teres, Pastides, Avrunin and Steingrub (1985) introduced the Mortality Probability Models (MPM) as an alternative to APACHE and SAPS. The uniqueness of MPM is that it does not require computation of an aggregate severity of illness score, as practised in APACHE and SAPS. Instead, MPM applied a direct statistical modelling approach that incorporates clinical variables in binary responses in a logistic regression model. The estimated risk of death for each patient is calculated using the MPM predictive equation.

Over the years, APACHE, SAPS and MPM have undergone several revisions, and have evolved from being simple systems to complicated models that involved the use of complex statistical methods. APACHE IV (Zimmerman, Kramer, McNair, & Malila, 2006), SAPS 3 Admission Score model (Metnitz et al., 2005; Moreno et al., 2005) and MPM_0-III admission model (Higgins et al., 2007) are the latest editions that were developed using large multi-centre data sets and advanced statistical methods. These models incorporate the severity of illness scoring component with a predictive component that is capable of predicting mortality outcomes of critically ill patients. Further analysis of the similarities, differences and evolution of APACHE, SAPS and MPM will be discussed in the next chapter.

1.2 Problem statement and motivation

ICU prognostic models are widely used in developed nations such as the United States, Europe and Australia. However, these models are not that popular in developing countries in South East Asia due to cost constraints, as well as, lack of resources and infrastructures. A search through the literature revealed no previous work in the area of prognostic modelling of intensive care unit outcomes in Malaysian ICUs. This is fairly expected because implementation of a prognostic model is an extremely costly affair. Most hospitals in Malaysia do not have automated patient monitoring systems and data

collection is still being manually performed. The increased complexity and extensive data collection process necessitate the use of automation and information technology for implementation of the latest models.

The Malaysian Registry of Intensive Care (MRIC) (formerly known as National Audit on Adult Intensive Care Units (NAICU)) is responsible for assessing the services and performances of selected government and private ICUs in the country. The participating ICUs are required to use SAPS II (Le Gall, Lemeshow, & Saulnier, 1993) severity of illness scores, which is an updated version of SAPS. The ICUs are then ranked according to their performances in terms of SAPS II scores and outcomes of the audits are officially declared in annual reports (Tong et al., 2012). SAPS II is chosen as a benchmark in the national audits due to its simplicity and because the parameters in SAPS II are easily available even in ICUs at the district level (C.C.Tan, personal communication, November 18, 2013). The predictive component of SAPS II is not used in the reporting of ICU performance in the national audits, and assessment of ICU performance is entirely based on SAPS II scores. The fact that performance is based on SAPS II could be an incentive for some ICUs to provide imperfect data.

As SAPS II was developed thirty years ago, there is a possibility that the model may no longer be valid for current application in Malaysian ICUs. ICU predictive models that were developed a long time ago are likely to deteriorate in performance over time and usually do not demonstrate good uniformity of fit when applied to a recent database (Kramer, 2005). Deterioration in the performance of these models is likely caused by factors such as changes in the baseline characteristics of ICU patients, use of specific therapeutic measures, or improvements in quality of care due to advances in medical technology and infrastructures over time (Moreno & Matos, 2000).

There is currently a lack of research in ICU prognostic modelling in Malaysia. In the author's opinion, the current assessment of ICU performance can be further

enhanced through implementation of a prognostic model that offers a patient stratification system, as well as, a mortality prediction component. In order to achieve a better understanding of the capabilities and performances of the latest prognostic models, the features and limitations of these models are reviewed and summarised in Chapter 2. A decision on the most suitable model to be used as a reference in this study is then made based on the analysis of features available in each model. The methodology employed in the reference model can then be examined and be used as a framework for development of a more suitable model that can be applied in the Malaysian context.

1.3 Scope of study

One of the major operational issues to be considered is to identify whether to focus the study in a single institution or multiple centres. Involvement of multiple centres will benefit and enhance the quality of this study, where findings that are obtained can be more meaningful, nationally representative and generaliseable. However, most of the government or private hospitals in Malaysia are facing under-staffing issues in their daily operations and some are not equipped with adequate facilities. The lack of response and commitment from these hospitals restricted the scope of this study to a single-centre ICU. Data that are obtained from a single-centre ICU is considered sufficient for this study since the research is focused on the modelling aspects instead of looking into performance comparisons between different ICUs.

1.4 Research questions

The study is divided into two stages. The first stage involves validation of an existing prognostic model in a Malaysian ICU, whereas the second stage involves development of a new prognostic model. This research aims to address the following questions:

Stage 1: Validation of an existing prognostic model in a Malaysian ICU.

i) Can any of the existing ICU prognostic models be adapted for application in a Malaysian ICU? Which model should be used for reference in the study?

ii) How well can the chosen model fit the Malaysian data? What can be interpreted from the results?

iii) What are the limitations of the chosen reference model? How can these limitations be addressed? Can the methodology be improved?

Stage 2: Development of a new prognostic model in a Malaysian ICU.

i) What alternative methodologies, other than a frequentist approach, can be used to develop a suitable prognostic model in the Malaysian ICU?

ii) How is the performance of models developed using alternative modelling strategies compared to a model developed using a frequentist approach?

iii) What is the most suitable model to be used in the Malaysian ICU?

1.5 Objectives of study

The following are the objectives for the two stages of study:

Stage 1: Validation of an existing prognostic model in a Malaysian ICU.

i) To identify and choose a suitable recent ICU prognostic model to be used for reference in a particular Malaysian ICU by performing a comprehensive review of existing well-established ICU prognostic models.

ii) To investigate the validity and accuracy of the chosen model in a Malaysian ICU by performing an external validation of the chosen reference model.

iii) To determine the limitations and gaps in the statistical methodology of the reference model, and identify areas for improvement.

Stage 2: Development of a new prognostic model in a Malaysian ICU.

i) To propose alternative techniques in the modelling of ICU mortality risk.

ii) To compare the performance of models developed using alternative modelling strategies against a model developed using a frequentist approach.

iii) To propose the best model for prediction of individual mortality risk in a Malaysian ICU.

In the first stage, the first objective of this study is to investigate the validity and accuracy of the chosen reference model in the single-centre Malaysian ICU. This involves conducting an external validation of the chosen model in order to determine its suitability and accuracy in the Malaysian ICU. There is a possibility that the chosen reference model in this study may not be suitable for application in a Malaysian ICU. The ability of a prognostic model to generalise for application in a different population is usually influenced by factors such as geographical location and methodological approach (Justice, Covinsky, & Berlin, 1999). Markgraf, Deutschinoff, Pientka, Scholten and Lorenz (2001) claimed that the prediction accuracy of prognostic models may not be applicable to external populations due to differences in case mix. Other potential factors that may affect the predictive accuracy of the models include lifestyle and cultural differences, ethnic and genetic dispositions, systematic differences in clinical practice, differences in measurement of physiological variables or medical definitions, as well as, the quality of medical services and treatment provided.

APACHE, SAPS and MPM are well-established systems that have evolved through several generations. However, despite being continually improved and revised over time, there are still some inherent limitations and inaccuracies in their statistical assumptions and methodologies. The predictive equations in these models were all built upon multiple logistic regression technique, where the maximum likelihood estimation (MLE) method was used for parameter estimation and variable selection. Although the MLE method is generally favourable for large and well-balanced data sets, it is not

appropriate for sparse data sets (Mehta & Patel, 1995) and tends to produce unreliable inferences when the number of model parameters is large relative to the size of data (Cox & Hinkley, 1974).

There is also an element of subjectivity in the assignment of points and their ranges for the physiological variables in APACHE and SAPS models. The approach in using worst physiological variables in these models is a subject of contention because it may not be the best representative of a patient's actual condition and may be affected by detection bias. This is because the choice of worst values is highly dependent on the measurement intervals for the physiological variables. Variability in the choice of worst values may occur due to differences in measurement frequencies for the affected variables. In most ICUs, the frequency of data collection for variables that are easily measured is often higher compared to variables that require laboratory analysis. Worst values are chosen based on available measurements, where unobserved variables are assumed normal. This assumption may affect the predictive accuracy of the prognostic models, resulting in underestimation of mortality risk (Holmes, Gregoire, & Russell, 2005). Furthermore, estimation of regression coefficients in these models was restricted to single-point estimation that was based on the worst physiological values within the first day of ICU admission.

It is evident that although mortality predictive models have been firmly established and revised over time, there are still some limitations in the existing models. With this in mind, the second part and main contribution of this study is to address the limitations and theoretical gaps in the existing models, by exploring more innovative and better alternative techniques in the modelling of ICU mortality outcomes. This corresponds to the second core objective of this study, which is to propose and develop a customised ICU prognostic model that is suitable for application in a Malaysian ICU.

1.6 Thesis outline

Chapter 1 briefly gives an overview of severity of illness scoring systems and ICU prognostic models, and an outline of the problem statement and motivation behind this study.

The literature review of this study is divided into two parts. The first part in Chapter 2 covers the literature review on the evolution of APACHE, SAPS and MPM systems, where the features, advantages, limitations and performances of the models are compared and discussed. The second part provides the literature review on the statistical methodologies that are being employed in this study.

Chapter 3 elaborates the scope and settings of this study, patient selection and exclusion criteria, as well as, variables and data that are being collected for the study. This chapter also explains the conceptual framework of the reference model that is chosen for the study and the methodology employed in validating this model in the Malaysian ICU. The methodology for construction of the proposed models in this study is specifically discussed in this chapter.

Chapter 4 presents the results and findings of this study. The first section of this chapter covers a detailed analysis on the demographic characteristics of patients included in the study. This is followed by an assessment of the performance of the reference model in this study. The third section of this chapter is focused on the performance comparison among the proposed models. The last section of this chapter reports the findings and results of an alternative method that was used to predict in-ICU mortality risk in this study.

The last chapter summarises and concludes the overall findings of this research. This chapter also includes discussion on some relevant issues, open problems, limitations and recommendations for future work.

CHAPTER 2: LITERATURE REVIEW

PART ONE: LITERATURE REVIEW ON INTENSIVE CARE UNIT SCORING SYSTEMS AND PROGNOSTIC MODELS

Over the years, there has been a rapid growth in the development of severity of illness scoring systems and prognostic models in critical care. Severity of illness scoring systems that are used in intensive care can be meant for specific or generic applications. Specific scoring systems are only applicable for certain types or groups of patients, whereas generic systems are used to evaluate almost all types of patients. These scoring systems can further be classified into three categories, i.e. anatomical, therapeutic and physiological. Anatomical scoring systems are used to assess the extent of injury and are useful for trauma audits and research, whereas therapeutic systems are used to quantify severity of illness among critical care patients based on the type and amount of treatment received (Gunning & Rowan, 1999).

The literature review for this study is focused on three generic physiological-based scoring systems, i.e. Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology (SAPS) and Mortality Probability Models (MPM). This chapter aims to provide an insight on the evolution of these systems over the years and to highlight the changes and improvements in each model revision. The features, advantages, limitations and performance of each model are also discussed in this chapter. A summary of the models that are covered in this chapter is shown in Table 2.1.

Table 2.1: Intensive Care Unit Prognostic Models that are included in literature review.

	Model	Year	Author	Origin
APACHE	APACHE	1981	Knaus et al. (1981)	U.S.
	APACHE II	1985	Knaus et al. (1985)	U.S.
	APACHE III	1991	Knaus et al. (1991)	U.S.
	APACHE IV	2006	Zimmerman et al. (2006)	U.S.
SAPS	SAPS	1984	Le Gall et al. (1984)	France
	SAPS II	1993	Le Gall, Lemeshow & Saulnier (1993)	Europe
	SAPS 3 admission	2005	Metnitz et al. (2005) ; Moreno et al. (2005)	Worldwide
MPM	MPM	1985	Lemeshow et al. (1985)	U.S.
	MPM II	1993	Lemeshow et al. (1993)	Europe
	MPM_0-III admission	2007	Higgins et al. (2007)	U.S., Canada, Brazil

2.1 Acute Physiology and Chronic Health Evaluation (APACHE)

2.1.1 APACHE

In early 1980s, Knaus et al. (1981) introduced the first generation of generic physiological scoring system, known as Acute Physiology and Chronic Health Evaluation (APACHE). APACHE was developed and validated using data from 805 consecutive eligible medical admissions to the George Washington University Medical Centre multi-disciplinary ICU in the United States. Burn and paediatric patients were referred to other hospitals and were excluded from the data set.

Development of APACHE was based on the premise that clinical factors such as patient's age, pre-existing health condition, physiological abnormalities and acute diagnoses can effectively estimate the risk of death of ICU patients (Holmes et al., 2005). APACHE was designed for use in the first day of stay in the ICU and captures patient data within the initial 32 hours of patient's stay in the ICU. This interval was chosen to allow ample time for all important patient data to be monitored and recorded

(Wagner et al., 1983). Missing data for variables that were not measured due to specific reasons were assumed normal.

APACHE consists of two main components, i.e. Acute Physiology Score (APS) and Chronic Health Evaluation (CHE). The first component quantifies severity of illness by measuring the patient's physiological abnormalities. The APS consists of a weighted sum of thirty-four physiological variables that were initially identified by a panel of clinicians to have potential influence on patient outcomes during ICU stay. These clinical and laboratory variables were derived from eight major organ-related categories (cardiovascular, respiratory, renal, gastrointestinal, haematologic, neurologic, metabolic and septic). The list of variables and points for physiological variables in APACHE is shown in Appendix A (Table A1). The selection and scoring of points for these physiological variables were done by a panel of ICU experts through clinical judgement. Points were assigned to the worst observations for each physiological variable within the first 32 hours following ICU admission. The majority of variables were individually assigned points between 0 and 4. Some of the variables were assigned scores between 0-1 and 0-2 points. Abnormal physiological observations were allotted higher points. The APS is computed by combining the points for all physiological variables. The range of APS falls between 0 and 129, where a higher value indicates greater severity of illness and a higher probability of mortality (Knaus et al., 1981). This scoring system offered an objective and quantitative approach to measure the severity of illness of a mixed group of adult patients who are severely ill, and to stratify them into different subgroups according to their associated risk categories.

The second component of the APACHE is the CHE (chronic health evaluation), which indicates the physiological reserve (age and existing chronic illnesses) of patients prior to ICU admission. Upon admission, patients are required to answer some questions pertaining to their health status, frequency of physician visits and daily

activities. Based on the responses given, the patients are classified into one of four categories (A, B, C and D), with A indicating good health condition and D being severely ill. Details of the questions and categories (extracted from the Medical Algorithms Project, 2008) are provided in Appendix A (Table A2). The APACHE score is derived by combining the APS and CHE category. In addition, APACHE required diagnosis of patients' primary type of disease as cause for admission (Knaus et al., 1981).

APACHE demonstrated superior accuracy in stratifying patients according to their risk categories and performed well in other countries such as France, Spain and Finland (Knaus, 2002). However, the system lacked probability calculations for the prediction of risk of death. Its application was complicated and demanding due to the huge number of variables to be collected, and the data collection window of the first 32 hours upon ICU admission was considered as too lengthy (Wong & Knaus, 1991). In addition, there was also substantial evidence pointing towards possible inaccuracies in the weighting of the neurologic abnormalities in APACHE scoring system (Wagner et al., 1983). Wagner and colleagues also highlighted that the APACHE classifications were not independent of therapy and were not appropriate for individual clinical predictions.

2.1.2 APACHE II

In order to address the limitations in APACHE, Knaus, Draper, Wagner and Zimmerman (1985) introduced APACHE II as a simplified version of its predecessor. This updated version was developed and validated using data from 1979 - 1982, based on 5815 ICU admissions in 13 hospitals in the United States. Similar to the original APACHE, the revised version consisted of three components; acute physiology variables, age and chronic health status. However, APACHE II established the concept

of a separate mortality predictive component that complements the existing severity of illness scoring component. Figure 2.1 illustrates the conceptual model of APACHE II, which forms the developmental framework for succeeding generations of APACHE models.

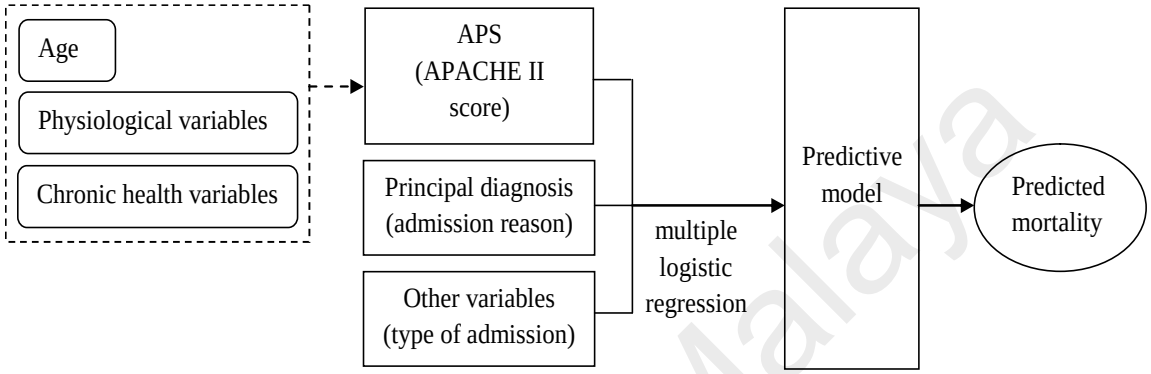


Figure 2.1: APACHE II conceptual model

Multiple logistic regression method was applied to develop the predictive equation, which incorporated variables such as the APACHE II score, type of admission and principal diagnostic categories (Knaus et al., 1985). The selection and weighting of physiological variables were still done through clinical judgement and careful evaluation of the role and impact of physiological measurements on patients' outcomes. The physiological variables defined in the original APACHE were reviewed, where the number of physiological variables was significantly reduced from thirty-four to only twelve in the updated version. These variables were removed based on clinical judgement. For instance, variables that were considered as unnecessary (e.g. blood urea nitrogen) or less frequently measured (e.g. serum osmolarity, serum lactate and skin anergy for testing) were excluded in the updated version (Wong & Knaus, 1991). The following physiological variables were retained in APACHE II: rectal temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation, arterial pH, serum

sodium, serum potassium, serum creatinine, haematocrit, white blood cell count and Glasgow Coma Scale (GCS) Score.

The list of physiological variables in APACHE II and their corresponding scores is shown in Appendix A (Table A3). APACHE II required mandatory data collection of all the twelve physiological variables, and assigned scores to the most abnormal readings throughout the first day after ICU admission. In APACHE II, the first day interval was shortened from the first 32 hours (original APACHE) to only the first 24 hours after ICU admission. Each of the physiological variables was allotted scores ranging from 0-4, except for GCS Score, which was assigned a range between 0 and 12. The GCS Score was given a higher weight since the measure of neurologic function was found to be underweighted in the previous version. The range for serum creatinine was revised in APACHE II, with acute renal failure being double-weighted. APACHE II also introduced assignment of scores for partial pressure of oxygen in arterial blood (PaO_2) values, for cases where the fraction of inspired oxygen (FiO_2) is defined as either $FiO_2 < 0.5$ or $FiO_2 \geq 0.5$ (Knaus et al., 1985).

Other than physiological variables, APACHE II required information such as patient's age, chronic health status and surgical status. Patients were classified according to their chronological age into one of five categories, with higher scores being assigned for older patients (see Appendix A, Table A4). This approach in categorising the age variable, which is continuous in nature, may lead to possible loss of power and statistical efficiency (Greenland, 1995). APACHE II adopted a different approach from its previous version in the evaluation of chronic health status. Instead of classifying patients into four different categories, scores were assigned to patients with existing severe organ system dysfunction based on their underlying disease. Non-operative or emergency surgery patients with underlying comorbidities were assigned a score of 5 points, whereas elective post-operative patients with immuno-compromised states were

given a score of 2 points. The combined total scores from patient's age, chronic health status and physiological components constitute the APACHE II score (Knaus et al., 1985).

APACHE II required identification of the main cause for ICU admission. Patients were assigned a unique main diagnosis to indicate cause for ICU admission based on an inventory of 50 diagnostic categories. These disease categories were classified according to major organ-related functions for medical (non-operative) and post-operative admissions. The diagnostic categories for non-operative and post-operative admissions are available in Appendix A (Tables A5 and A6) respectively. Patients who do not fall under any of the specific diagnostic classification were assigned in one of the five general categories (metabolic/renal, respiratory, neurologic, cardiovascular and gastrointestinal).

The predicted mortality rate for groups of critically ill patients is given by the following equation, which includes the APACHE II score, presence of emergency surgery and disease category:

$$ln\frac{R}{1-R} = \text{-3.517 + (APACHE II score } \times 0.146) + (0.603, \text{ if postemergency surgery}) + (\text{Diagnostic category weight}). \quad (2.1)$$

The developers of APACHE II excluded post-coronary artery bypass surgery (CABG) patients in their patient sample due to significant differences in implications of CABG physiologic derangement compared to other types of ICU admissions. The inclusion of CABG patients would likely affect the model's predicted accuracy because these patients are often associated with low mortality risks despite having high initial severity of illness scores (Knaus et al., 1985).

Knaus et al. (1985) highlighted some of the limitations of APACHE II and gave some recommendations for future improvements. Firstly, they were of the opinion that

data collection should also include admission values, as they observed that most of the abnormal measurements within the first 24 hours after ICU admission were close to admission values. The approach in using admission values would have made the severity classification to be independent of therapy. APACHE II was also criticised for its failure to accommodate several important factors in its predictive equation. Dragsted et al. (1989) argued that the significance of lead time bias should be taken into account, where lead time bias is defined as the different lengths of time that patients are ill prior to acute illness (Holmes et al., 2005). In addition, Escarce and Kelley (1990) proposed that factors such as treatment received and location of patients prior to ICU admission should be included in the APACHE II equation. These factors were considered important as they may cause changes in the physiological variables and influence the APS score. The requirement to assign patients to only one diagnostic category based on the principal reason for ICU admission was also considered to be restrictive and may introduce bias (Cowen & Kelley, 1994). For example, it would be difficult to make a decision regarding a patient with multiple symptoms. Improper assignment of diagnosis will eventually affect the accuracy of predicted mortality.

Other than being employed as a tool for comparing quality assurance in ICUs, APACHE II was also used to risk-stratify patients in order to control case mix, so that appropriate comparisons of therapy could be performed (Wong & Knaus, 1991). Although APACHE II was developed in the United States, it was successfully validated in other countries such as New Zealand (Zimmerman et al., 1988), Japan (Sirio et al., 1992) and Canada (Wong, Crofts, Gomez, McGuire, & Byrick, 1995). These studies observed that despite differences in case mix and medical practices, the APACHE II score was fairly accurate and reliable in predicting group mortality outcomes in their respective populations. In studies to evaluate the accuracy of APACHE II in mortality prediction against clinical assessment, the performance of APACHE II was found to be

comparable, if not, superior to clinical judgement (McClish & Powell, 1989; Silverstein, 1988).

On the other hand, a large multicentre study by the UK Intensive Care Society reported contradictory results, where the APACHE II equation was found to be inappropriate for the UK data due to differences in clinical definitions and interpretation (Rowan et al., 1994). The investigators compared the outcomes among 26 ICUs in Britain and Ireland both before and after adjustment for case mix using APACHE II. The overall goodness-of-fit of the APACHE II equation for the 26 ICUs was good, but poor uniformity of fit was observed when patients were grouped by age, diagnosis or APACHE II score. A separate study by Goldhill & Withington (1996) also reported the failure of APACHE II to accurately adjust for case mix in 19 ICUs in the UK.

2.1.3 APACHE III

In 1991, the third instalment of APACHE became commercialised with the introduction of APACHE III (Knaus et al., 1991). This version, which was introduced as a proprietary database and decision support system, was originally distributed by APACHE Medical Systems (McLean VA). It is now currently being managed by Cerner Corporation (Kansas City, Missouri, USA). Similar to APACHE II, the updated version consisted of two components; the APACHE III score and the mortality predictive equation. However, APACHE III was developed based on a much larger patient database, comprising 17,440 adult medical/surgical ICU admissions in 40 US hospitals.

Data collection for this study was conducted for 1.5 years, starting from May 1998 until November 1989. In the original APACHE, only paediatric and burn patients were excluded from analysis, whereas APACHE II excluded post-operative coronary artery bypass graft patients. In APACHE III, the following patients were excluded:

those who were less than 16 years old, burn patients, and acute myocardial infarction patients. Patients with less than 4 hours of ICU stay were also removed from the study. APACHE III included data collection for post-coronary artery bypass graft (CABG) patients. However, data for these patients were reported separately as APACHE III offered different predictive equations according to the type of admission, i.e. non-CABG model and CABG model (Becker et al., 1995).

In the non-CABG model, the total number of physiological variables was increased from 12 (APACHE II) to 17 (APACHE III). These physiological variables were selected through a combination of clinical judgement and statistical assessment. Initially, 20 potential physiological variables were shortlisted to be important predictors of mortality through previous experience and clinical judgement. Multivariable logistic regression approach was then applied to determine the relationship and interactions between mortality rate and each of the 20 variables. This method was also used to derive the ranges and scores for the physiological variables. All of the variables in APACHE II were retained except for serum bicarbonate and serum potassium. These two variables were removed because they were found to be not statistically significant. Six new variables were included in APACHE III, i.e. blood urea nitrogen (BUN), urine output, serum albumin, bilirubin, glucose and a combined variable (serum pH and pCO_2) for acid-based abnormalities (Knaus et al., 1991). The ranges and scores for all of the physiological variables were completely redefined in APACHE III. Each of the variables was divided into several clinical ranges and scores were assigned to each of the categories, with one being the normal category. Scores were only assigned to the worst physiological measurements observed within the first 24 hours of ICU admission. Patients with worst values that fall within the normal range were not assigned any scores. Higher scores were allocated for worst physiological measurements that deviate further from the normal category. Assessment of Glasgow Coma Scale (GCS) score was

also modified in APACHE III to improve the accuracy of assessment of neurologic function. Instead of individual scores for each of the components required to assess GCS, scores were assigned to various combinations of eye, verbal and motor components (Knaus et al., 1991). The complete list of physiological variables in APACHE III is shown in Appendix A (Table A7). Physiological values that were not recorded were assumed normal and not given any score.

Retaining the concept used in APACHE II, APACHE III score is the sum of points for age, chronic health status and worst physiological observations within the first 24 hours of patient's stay in the ICU. However, the points and ranges for chronic health and age variables were modified in APACHE III. The age variable was divided into more categories, where higher allocation of scores was given to older patients. For instance, patients older than 85 years old were assigned a score of 24 points in APACHE III, whereas similar patients were only assigned a score of 6 points in APACHE II. Evaluation of chronic health status involved assignment of scores to seven comorbidities; acquired immunodeficiency syndrome (AIDS), hepatic failure, lymphoma, metastatic cancer, leukaemia/multiple myeloma, immunosuppression and cirrhosis. Additional scores were given to non-operative or emergency surgery patients with underlying comorbidities. Modification of scores to these variables resulted in a higher variability in the overall APACHE III score, ranging between 0 and 299 (Knaus et al., 1991). The detailed allocation of scores for age and comorbidities in APACHE III is shown in Appendix A (Table A8). To improve accuracy in disease identification, the total principal diagnostic groups was further extended from 50 (APACHE II) to 78 (APACHE III). The list of disease categories is given in Appendix A (Table A9). The coefficients of these diagnostic categories were not available in public domain. APACHE III addressed the shortcomings in APACHE II by including variables to account for patient's source and treatment obtained before ICU admission, and the

difference in duration between emergency room and ICU admission (Knaus et al., 1991). APACHE III offered daily predictions of hospital mortality for individual patients, by providing predictive equations for the first seven days of stay in the ICU (Wagner, Knaus, Harrell, Zimmerman, & Watts, 1994). In addition, separate predictive equations for patients who had coronary artery bypass graft (CABG) surgery were also provided in APACHE III (Becker et al., 1995).

Knaus et al. (1991) discussed three potential advantages of using APACHE III over clinical judgement. Firstly, since the prognostic estimates were derived from reproducible data, they should be more reliable compared to individual judgement. Next, APACHE III was built using a large reference database and should be more representative of the population of interest. Predictions in APACHE III also reflected the patient's response to treatment, irrespective of the order in which the patient was admitted into ICU. However, the APACHE III developers cautioned that the APACHE III score is only suitable to be used independently to stratify patients according to their severity of illness, within homogeneous disease categories.

On the whole, APACHE III demonstrated good calibration and discrimination, with an area under the receiver operating characteristic curve value of 0.90 and a total correct classification rate at 50% mortality risk level of 88% (Knaus et al., 1991). Although APACHE III had good discrimination, the model exhibited poor calibration in several external validation studies. These findings suggested that APACHE III might not be suitable for use in other countries or populations with different characteristics. In their study which involved 10 Brazilian ICUs, Bastos, Sun, Wagner, Knaus and Zimmerman (1996) found that APACHE III provided good discrimination, despite a high overall standardised mortality ratio (SMR). However, APACHE III exhibited poor calibration and uniformity of fit in the Brazilian study.

Three years later, Pappachan, Millar, Bennett and Smith (1999) conducted the largest assessment of APACHE III in the United Kingdom, involving 12,793 patients admitted to 17 ICUs in the South of England from 1 April 1993 to 31 December 1995. The study revealed that the observed overall hospital mortality for UK ICU patients was 25% higher than predicted, with an SMR of 1.25. Two possible explanations were given for this discrepancy. Firstly, the performance of the UK ICUs could in reality be poorer compared to US ICUs due to differences in the structure and organisation of intensive care, availability of technology and training resources between the two countries. The alternative explanation for the excess in observed mortality could be failure of the APACHE III equation to fit the UK data. Pappachan et al. (1999) argued that the second reason was more plausible since APACHE III was applied to a population with different composition.

In another study involving a German interdisciplinary intensive care unit, Markgraf, Deutschinoff, Pientka and Scholten (2000) found APACHE III to have insufficient calibration because the observed mortality rate was higher than predicted. The study suggested that differences in the patient selection and case mix, admission policies and lead time were potential factors that influenced the performance of APACHE III in the German cohort of patients. Moreover, Markgraf et al. (2000) believed that inaccuracies of mortality prediction for different subgroups and length of hospital stay were other factors that affected the accuracy of mortality prediction.

On the other hand, Cook et al. (2002) reported positive findings in an independent validation study in an Australia ICU. The study at the Princess Alexandra Hospital was based on 5681 consecutive eligible admissions from 1 January 1995 to 1 January 2000. APACHE III was found to have excellent discrimination and good calibration in their patient sample, despite differences in case mix between the Australian and APACHE III data sets. In another single centre study in South Korea,

Jeong, Kim and Kim (2003) also found that APACHE III exhibited good discrimination, calibration and uniformity of fit in a cohort of 284 patients. However, it is important to note that these positive results were obtained in single-centre settings, whereas the earlier validation studies by Bastos et al. (1996) and Pappachan et al. (1999) involved large multi-centre settings.

Zimmerman et al. (1998) performed an independent study to assess the accuracy and validity of APACHE III in 285 ICUs in 161 US hospitals from 1993 to 1996. The study, which was based on 37,668 ICU admissions, revealed that APACHE III exhibited excellent discrimination with an area under the receiver operating characteristic (ROC) curve value of 0.89. The results of the study indicated no significant difference between the aggregate observed and predicted hospital mortality. However, they observed that calibration was not perfect, as there were differences between observed and predicted hospital mortality across deciles of patient risk. In order to improve the accuracy of the APACHE III predictive equations, they suggested improvements in the precision of disease labelling, better acquisition and weighting of neurologic abnormalities, as well as, adjustment of coefficients to reflect a larger database and differences in treatment outcomes over the past five years.

2.1.4 APACHE IV

APACHE IV was officially launched in 2006 as the successor to APACHE III. The model was developed from a more contemporary database, which consisted of 110,558 ICU admissions at 104 ICUs/coronary care units in 45 US hospitals. Data collection for this multicentre study started from 1st January 2002 until 31st December 2003. The study excluded patients who had less than 4 hours of ICU stay, those under age 16 years old, burn victims and those who were admitted after transplant operations (except for hepatic and renal transplants). The study also ruled out patients with missing APS on

the first day of ICU admission, and those who were hospitalised for more than 365 days. Patients who were transferred from another ICU were also not considered so as to eliminate potential biases that may be caused by any clinical interventions or therapy received prior to ICU admission (Zimmerman et al., 2006).

Development of APACHE IV focused on a complete revision and revalidation of all existing APACHE III equations. More than half of the APACHE III equations required remodelling, while some equations were eliminated as they were no longer relevant to current needs. Multivariable logistic regression technique was employed in the remodelling process of APACHE IV equations. Modelling of APACHE IV equations was done through two statistical methods; logistic regression for dichotomous outcomes and linear regression for continuous outcomes. The regression splines approach by Stone and Koo (1985) was used in APACHE IV to improve model precision. Restricted cubic regression splines were applied to enable age, pre-ICU length of stay and APS predictors to have flexible non-linear relationship with the response variable. The measurement unit for pre-ICU length of stay was also changed from integer to continuous scale.

Data items that were collected in APACHE IV were similar to those in APACHE III. These included common variables such as age, physiologic data and chronic health conditions. All of the physiological variables in APACHE III were retained in this updated version, with no changes being made to their scores (see Appendix A (Table A7)). However, an important change was made to the computation of the severity of illness score. The APACHE III severity of illness score was computed by combining the scores for three components, i.e. age, physiology and chronic health variables. In APACHE IV, the severity of illness score is known as Acute Physiology Score (APS). The APS is computed by considering the physiology component alone, by combining the scores for all of the worst physiological variables within the first day

after ICU admission. The range of APS was between 0 and 252. Chronic health and age variables were excluded in the APS calculation, but directly incorporated into the APACHE IV predictive equation (Zimmerman et al., 2006).

APACHE IV provided separate predictive equations for coronary artery bypass graft (CABG) patients and non-CABG patients. Some of the variables to be collected for non-CABG patients included admission source, emergency surgery status and pre-ICU length of stay. Four new clinical variables were included into the predictive equation, i.e. mechanical ventilation status, failure to obtain Glasgow Coma Scale (GCS) score owing to patient being under sedation or paralysis, application of thrombolytic therapy and a rescaled PaO_2 to FiO_2 ratio. The complete list of variables to be collected for non-CABG patients is shown in Appendix A (Table A10). The comorbidities in APACHE III were maintained in APACHE IV. However, the number of disease categories for non-operative and post-operative admissions was further increased to 116 (see Appendix A (Table A11)). The new categories were chosen based on factors such as frequency, clinical homogeneity and the impact of each diagnostic category on mortality rate (Zimmerman et al., 2006).

The data items to be collected for CABG admissions are summarised in Appendix A (Table A12). Variables to be recorded for CABG patients included cardiac related information such as the number of grafts, prior CABG surgery, whether the type of graft is internal mammary, and myocardial infarction during current hospitalisation. Other variables that were collected for CABG patients were gender, pre-ICU length of stay, emergency surgery status and diabetes.

Some changes were made to the daily mortality prediction equations in APACHE IV. Two new variables were added into the equations for Day 2 and Day 3-8. The equation for daily prediction of mortality is given as:

$$\text{daily risk} = \begin{pmatrix} \text{APACHE day 1} \\ \text{prediction} \end{pmatrix} + \begin{pmatrix} \text{influence of} \\ \text{APS current day} \end{pmatrix} + \begin{pmatrix} \text{influence of change} \\ \text{in APS since yesterday} \end{pmatrix}. \quad (2.2)$$

Zimmerman and Kramer (2008) explained that difficulties in calculation of daily mortality estimates arise mainly due to complexity of patient data itself. They advocated the use of an electronic interface to address the complexity of data collection in APACHE IV. Other than being used for prediction of group mortality, APACHE IV was also designed for prediction of group resource use, i.e. prediction of ICU length of stay and prediction of risk for patients receiving active therapy (Zimmerman & Kramer, 2008).

Overall, Zimmerman et al. (2006) found that APACHE IV exhibited excellent discrimination (area under receiver operating characteristic curve value of 0.88) and good calibration, despite a large validation sample size. They claimed that APACHE IV can be used to benchmark performance of ICUs in the US, to evaluate quality of care and perform disease-specific subgroup analyses. Some of the factors that contributed to the accuracy of APACHE IV included its successful use of physiologic abnormalities for risk adjustment, rescaling of PaO_2/FiO_2 and Glasgow Coma Scale variables, improved precision of disease labelling and the use of splines for age, APS and prior length of stay variables. However, the developers of APACHE IV believed that APACHE IV might not be suitable to be applied internationally due to differences in infrastructure, managerial policies and quality in patient care. They also cautioned that the model's accuracy was likely to deteriorate over time due to future changes in clinical policies and practices.

The performance of APACHE IV was externally evaluated in single-centre settings in Saudi Arabia (Kherallah et al., 2008), India (Bhattacharyya & Todi, 2009; Parajuli, Shrestha, Pradhan, & Amatya, 2015) and South Korea (Jae et al., 2017). Positive results from these studies suggested the possibility of APACHE IV being

robust enough for application in other countries and in single-centre settings, even though it was developed based on data from multiple institutions.

In a larger study involving 44,112 patients in 59 mixed medical-surgical Dutch ICUs, Brinkman et al. (2011) performed an external validation on APACHE IV and compared its performance with APACHE II and SAPS II. APACHE IV was found to have good accuracy and discrimination (area under receiver operating characteristic curve value = 0.89), but poor calibration ($\hat{C}$ statistic = 822.67). Further customisation was performed on APACHE IV and improvement in calibration was observed ($\hat{C}$ statistic = 142.32). The older models APACHE II and SAPS II were found to be not suitable for their current study and required customisations.

There were also several studies that evaluated the performance of APACHE IV in specific subgroups of patients. Lin et al. (2007) found the APACHE IV to be excellent in providing short term prognosis for critically ill patients who were receiving extracorporeal membrane oxygenation (ECMO) in a specialised intensive care unit in Taiwan. In another study, Costa e Silva et al. (2011) conducted a comparison of APACHE IV, SAPS 3 Admission Score and MPM-III models in acute kidney injury critically ill patients who were admitted to six ICUs in a teaching tertiary care centre in Brazil. The study concluded that although all of the three models had almost similar discriminatory abilities, MPM-III performed the worst, in terms of calibration. The customised regional equation of SAPS 3 presented the best fit for the cohort of patients, whereas APACHE IV underestimated mortality despite having acceptable calibration.

In a single-centre study in China, Xing et al. (2015) compared the performance of APACHE II, SAPS 3 Admission Score Model and APACHE IV in critically ill cancer patients in a single tertiary hospital. Although discrimination and calibration were good in all three models, SAPS 3 and APACHE IV underestimated in-hospital mortality, whereas APACHE II overestimated in-hospital mortality. The study also

concluded that the overall performance of SAPS 3 was superior to APACHE II and APACHE IV.

On the other hand, a study by Nassar Junior et al. (2013) revealed that APACHE IV performed better than SAPS 3 in a cohort of 1,065 acute coronary syndrome patients in three Brazilian ICUs. SAPS 3 was reported to have inadequate calibration for the specific group of patients, whereas APACHE IV exhibited adequate calibration and good discrimination. In a different study, Hu et al. (2013) compared the predictive accuracy of APACHE IV against MELD (Model for End-Stage Liver Disease) scoring system for patients who were admitted to a single centre ICU in China, after orthotopic liver transplantation. APACHE IV was found to have better discrimination than MELD, although both systems appeared to be well-calibrated.

2.1.5 Comparison of APACHE models

A summary of the evolution and differences of the APACHE models is given in Appendix A (Table A13). The APACHE models were all developed and validated in the United States. With the exception of the first generation, the rest of the models were developed using data from multiple institutions. All of the four APACHE models were scoring systems based on physiologic data obtained from patients who were admitted to the ICU, and shared three common components, i.e. age, physiological and chronic health variables. APACHE II remains the simplest version until now, with the least number of physiological variables among the four versions. The selection and weighting of variables, which were initially done by clinical judgement in APACHE and APACHE II, were complemented by multiple logistic regression approach in the subsequent models. All models require the identification of a principal admission diagnostic category for patients who are admitted to the ICU. With each revision, the number of diagnostic categories was increased in order to improve precision in disease

labelling. Although this contributed towards making the model more complicated, recent advances in computer technology have made the process of data collection easier through application of electronic interfaces.

Development of newer models such as APACHE III and APACHE IV involved the use of larger patient databases and more advanced statistical methods. The introduction of new variables and application of regression splines in APACHE IV have made this model to be the most complicated among the four versions. APACHE III and APACHE IV provided separate predictive models for postcoronary artery bypass graft patients and offered daily risk predictions of ICU and hospital mortality. These two models also predicted risk of therapy for individual patients and remaining ICU length of stay for patients who are still in the ICU after 5 days.

Past studies have suggested deterioration in the accuracy of the older models over time. Zimmerman et al. (2006) argued that APACHE II mortality predictions are not likely to be accurate when used in large contemporary databases due to the absence of multiple predictor variables. They further recommended that APACHE III be used only as a summary measure of severity of illness, but not for comparing observed and predicted outcomes. On the other hand, the cost associated with the purchase of software constitutes a major limitation in the utility of APACHE III. As the latest version, APACHE IV has undergone many enhancements and refinements that addressed the shortcomings of the earlier versions. As such, APACHE IV remains useful as a suitable reference database for benchmarking purposes until it is replaced by a newer model.

2.2 Simplified Acute Physiology Score (SAPS)

2.2.1 SAPS

Le Gall et al. (1984) introduced Simplified Acute Physiology Score (SAPS) in 1984 as a simplified version of the first generation of APACHE. The model was developed based on 679 consecutive admissions in 8 French ICUs. Paediatric and burn patients were excluded in the development of the model. Based on the list of 34 physiological variables in the original APACHE version, only 13 physiological variables and age variable were considered as significant for inclusion in the first version of SAPS. The selection and assignment of weights to variables were done through clinical judgement. Each variable was assigned scores from 0 to 4 for the worst physiological measurement collected within the first 24 hours after ICU admission. The list of physiological variables in SAPS is given in Appendix B (Table B1), while the scoring for age variable is displayed in Appendix B (Table B2).

The severity of illness score in SAPS is known as the SAPS score. This score is calculated by combining the scores for age variable and worst physiological variables within the first 24 hours after ICU admission. As compared to the first version of APACHE, the chronic health status component is not included in the first version of SAPS. The total SAPS score can assume any value between 0 and 56 points and is less than the APACHE score. This score is useful in providing risk stratification for critically ill patients based on their prognosis. Although it did not have an integrated mortality prediction component to estimate mortality rate for individual patients, the developers of SAPS provided a general conversion table that relates SAPS scores to mortality rates (see Appendix B (Table B3)). Specific mortality rates for different SAPS scores were also given according to the type of ICU admission (see Appendix B (Table B4)). These estimates were obtained through empirical means (Lemeshow, Teres, Avrunin, & Pastides, 1987).

2.2.2 SAPS II

Le Gall et al. (1993) introduced the second generation of Simplified Acute Physiology Score (SAPS II). This European/North American multi-centre study involved a large sample of 13,152 consecutive medical and surgical admissions between 30 September 1991 to 28 February 1992 from 137 ICUs in 12 countries. The development process involved 65% of the total available patients, while the rest of the patients were chosen for the validation data set. The study excluded data for patients who were under 18 years of age, burn patients, coronary care patients and cardiac surgery patients. Those with missing data on type of admission and ventilation status were also excluded.

In terms of data collection, SAPS II required similar types of information as in APACHE II, such as patient demographic details, physiological variables, age, chronic health diagnoses and types of admission (scheduled surgical, unscheduled surgical or medical). Initially, the developers of SAPS II identified 37 potential variables, which comprised all of the variables in SAPS and some additional new variables that were thought to have potential influence on mortality risk. Using multiple logistic regression, they found that only 17 variables (12 physiological, age, type of admission and 3 comorbidities) were significant to be included in the SAPS II model. Two physiological variables in SAPS were removed (serum glucose and haematocrit) and one new variable was included (bilirubin) in the updated version. AIDS, haematologic malignancy and metastatic cancer were identified as important chronic health variables. For this version, the PaO_2/FiO_2 ratio was not recorded for patients who were not ventilated or receiving continuous positive airway pressure (CPAP).

Scores that were assigned to all the variables in SAPS II were obtained through logistic regression modelling. Similar to APACHE II and SAPS, SAPS II used the worst physiological data recorded within the first 24 hours of ICU admission. Scores were also allocated for the type of admission and chronic health status. Patients who were

admitted for unscheduled surgery were assigned higher scores compared to those who were medical admissions. Missing data that were not available were assumed to be within normal limits. Appendix B (Table B5) provides a summary of the physiological variables in SAPS II and their weightings. Scores for other variables in SAPS II are shown in Appendix B (Table B6).

The SAPS II score is obtained by summing up all the scores for the 17 variables. The range of this score lies between 0 and 160. Unlike its predecessor, SAPS II offered prediction of hospital mortality through a multiple logistic regression equation. The probability of hospital mortality is computed as

$$P(\text{hospital mortality}) = \frac{e^{logit}}{1+e^{logit}}, \tag{2.3}$$

where the *logit* term defined in Le Gall et al (1993) is given as

$$logit = -7.7631 + 0.0737 \times (\text{SAPS II score}) + 0.9971\left[ln\,(\text{SAPS II score} + 1)\right]. \tag{2.4}$$

Compared to the APACHE models, SAPS II did not require the selection of a primary admission diagnosis. The rationale for excluding this variable in the model was due to difficulties in categorising patients with multiple diagnoses into a category (Le Gall et al., 1993).

The performance of SAPS II was found to be significantly superior to its predecessor. SAPS II demonstrated excellent discrimination, with area under the receiver operating characteristic curve values of 0.88 (developmental data set) and 0.86 (validation data set). The model also indicated good calibration, with large *p*-values obtained in the goodness-of-fit tests performed on the developmental (*p*-value = 0.883) and validation (*p*-value = 0.104) data sets. In a study to compare the performances of SAPS and SAPS II in Italian ICUs, Bertolini et al. (1998) concluded that the performance of SAPS II appeared to be better than SAPS. Although both models did not fit the Italian cohort of patients, SAPS II was found to have better discriminative ability and offered more accurate predictions compared to SAPS. The lack of fit of SAPS in the

Italian database was expected because the model was developed and calibrated solely based on French population. As for SAPS II, it is believed that problems such as patient selection and case mix, as well as, structural and organisation factors influenced predictive performance of SAPS II in their study.

The performance of SAPS II was compared to APACHE II in several validation studies involving other countries. In a multicentre study in Portugal, Moreno and Morais (1997) concluded that SAPS II performed better than APACHE II, although both models demonstrated poor overall calibration. Likewise, Katsaragakis et al. (2000) obtained similar findings, with both SAPS II and APACHE II showing underprediction of mortality in their single-centre Greece study. Although the Greek investigators found that SAPS II had better discrimination, APACHE II performed better in terms of calibration.

Capuzzo, Moreno & Le Gall (2008) provided some explanations for the lack of fit of SAPS II in new populations. First, they argued that differences in case mix contributed to the dismal performance shown by SAPS II in new populations that were different from the original in which the model was developed. Secondly, improvement in medical care was also another possible factor that could have influenced the calibration of SAPS II over time. They suggested that the poor performance of SAPS II in new populations could also imply differences among ICUs in terms of quality of care.

Several investigators proposed customisation of SAPS II as a solution to improve model accuracy. In general, there are two types of customisation that can be applied in a predictive model. The 'first-level customisation' approach involves only adjustment on the severity of illness score itself, through computation of a new *logit* formula. This approach requires calculation of a new prediction equation without making any changes to the variables and weightings used in the original model (Metnitz et al., 2005). On the other hand, the 'second-level customisation' is considered more

complicated as it requires re-evaluation of each component of the score, or even addition of new variables (Capuzzo et al., 2008).

Metnitz et al. (1999) used the first-level customisation approach to propose a new customised model, known as SAPS II-AM, after a validation study indicated that SAPS II was not adequately calibrated when applied to a cohort of Austrian patients. The customisation was achieved through derivation of a new logistic regression equation. The new customised model improved the accuracy of prediction and demonstrated excellent goodness-of-fit. A year later, Metnitz, Lang, Vesely, Valentin and Le Gall (2000) performed a validation of SAPS II and the customised SAPS II-AM models in a large cohort of Austrian intensive care patients. Consistent with the results obtained in the earlier study, they observed poor calibration in SAPS II, especially when the patients were grouped according to the types of admission (medical, scheduled surgical and unscheduled surgical). SAPS II underestimated mortality in the lower risk deciles but overestimated mortality in higher risk deciles. Reasons for the poor performance of SAPS II were attributed to its inability to take into account all the factors that significantly influenced outcome and the lack of important variables that were not included in the model. On the other hand, the customised model of SAPS II-AM exhibited improved calibration and better fit in subgroups of scheduled and unscheduled surgical admissions. Metnitz and colleagues also proposed a new model, known as SAPS II-AM2, which was developed based on a larger and more recent Austrian database. The new model outperformed both SAPS II and the customised SAPS II-AM models, and demonstrated excellent calibration when patients were categorised by types of admission (Metnitz et al., 2000).

Aegerter et al. (2005) employed a second level customisation approach on SAPS II using retrospective data from 33,471 French patients admitted to 32 ICUs of the Cub-Rea group. This approach required re-evaluation of components and addition of new

variables in SAPS II. The customised model improved the uniformity of fit for different categories of patients except for diagnosis related groups. Although customisation helped to improve calibration and uniformity of fit, variations in case mix between data sets restricted comparisons of quality of care.

Le Gall et al. (2005) conducted a study from 1 January 1998 to 31 December 1999. using 77,490 admissions in 106 French ICUs and proposed an updated mortality prediction model of SAPS II, known as the expanded SAPS II model. Six new admission variables were included in the expanded SAPS II model, where the variables were sex, length of previous ICU stay, patient location prior to ICU admission, clinical category and whether drug overdose was present. These variables were chosen because they were easily obtainable and routinely measured during ICU admission. In addition, the scores for the age variable were also completely revised in the updated SAPS II version. Appendix B (Table B7) provides the list of additional variables and their respective weightings in the updated version, known as the expanded SAPS II model.

The performance of the expanded SAPS II model was compared against the original SAPS II and another customised SAPS II model. The original SAPS II model overestimated mortality and had poor calibration, despite showing good discriminatory power. The customised SAPS II model performed slightly better than the original version, with improved calibration. However, poor uniformity of fit was observed using this customised model. The expanded SAPS II model outperformed both the original and customised models, by demonstrating excellent calibration, better discrimination and good uniformity of fit. Although the expanded SAPS II model exhibited good fit in the validation set, there were still doubts as to its performance when applied to a different population from which it was developed (Le Gall et al., 2005).

Teres and Lemeshow (1999) highlighted that although most researchers believed that customisation or recalibration was the solution when the original model failed to

display good calibration, this assumption may be incorrect. They further explained that advances in medical science could have contributed towards improvement in quality of care over time. Although recalibration and customisation were able to improve the prediction accuracy to some extent, they were still not able to solve the inherent problems in these models, such as shifts in the baseline characteristics of the populations over time and lack of important prognostic variables (Metnitz et al., 2005).

2.2.3 SAPS 3 Admission Score Model

As SAPS II was found to be outdated, SAPS 3 Outcomes Research Group initiated the development of an updated model known as SAPS 3 Admission Score model to address the issues and limitations in SAPS II (Metnitz et al., 2005; Moreno et al., 2005). The study used data from 16,784 consecutive admissions in 303 ICUs from 35 countries worldwide. This multinational cohort study involved the participation of 7 different geographic regions; Australasia, Central and South America, Central and Western Europe, Eastern Europe, North America, Northern and Southern Europe, and Mediterranean countries (Metnitz et al., 2005). To date, this study remains the largest prospective multicentre and multinational study among all the available prognostic systems in intensive care.

The study excluded patients who were less than 16 years old, and only considered the first admission for patients with multiple admissions. Besides that, patients without ICU admission or discharge data, with more than 90 days of ICU stay, and those without 'ICU outcome' date were also excluded. In contrast to SAPS II, SAPS 3 Admission Score model included data for coronary care and cardiac surgery patients.

SAPS 3 Admission Score model required data collection during ICU admission, on days 1, 2 and 3, and on the last day of ICU stay. The information to be collected upon ICU admission included sociodemographic data, condition of patient before ICU

admission (chronic conditions and medical diseases), data about patient's condition at ICU admission (reason for admission, infection at admission, surgical status) and data about patient's physiologic abnormalities at ICU admission. On the other hand, data that was collected on days 1, 2 and 3, and the last day of ICU stay comprised severity of illness, length of ICU and hospital stay, and outcome data. In addition, the number and severity of organ dysfunction was also measured using Sequential Organ Failure Assessment (SOFA), which was described in Vincent et al. (1998). Other features of SAPS 3 Admission Score model include its ability to estimate patient's vital status at 28 days after ICU admission (Moreno et al., 2008) and investigate variability in outcome and resource use between ICUs (Rothen et al., 2007).

The data collection interval for SAPS 3 Admission Score model differed significantly from the earlier SAPS and APACHE models. Instead of collecting data within the first 24 hours after ICU admission, the admission data for SAPS 3 Admission Score model were collected within an hour before or after ICU admission. The rationale for limiting data acquisition within the first hour of ICU was to minimise the impact of overestimating mortality and enable prediction of mortality to be done before any ICU interventions (Moreno et al., 2005). The SAPS 3 investigators explained that this was a major advantage of the model. Other versions such as SAPS II, APACHE II and APACHE III were subject to influence by Boyd and Grounds effect (Boyd & Grounds, 1994), in which occurrence of more abnormal physiologic values during the first 24 hours after ICU admission may lead to an increase in computed severity of illness and predicted mortality.

There were more variables in SAPS 3 Admission Score model compared to SAPS and SAPS II. A total of 20 variables were included in SAPS 3 Admission Score model, based on a combination of expert judgement and logistic regression method. These variables were classified into three categories; Box I, Box II and Box III. The

first box consisted of 5 variables that represent patient characteristics before ICU admission. Box II consisted of 5 variables that were related to the circumstances of ICU admission, while Box III contained 10 physiological variables that were measured within 1 hour before or after ICU admission. Details of the variables are given in Appendix B (Tables B8, B9 and B10). Missing data were assumed as normal in this model.

The developers employed LOWESS (Locally Weighted Scatterplot Smoothing) (Cleveland, 1979) approach to group continuous predictive variables into mutually exclusive categories. Multidimensional tables and regression trees were also used together with clinical judgement to form classes of categorical variables. In their efforts to reduce model complexity, the SAPS 3 investigators applied stepwise logistic regression to identify significant predictors and interactions among predictors. Bootstrapping method was applied to check the stability of variable selection and reduce complexity processes (Moreno et al., 2005).

SAPS 3 Admission Score model provided a general mortality predictive equation for global use, as well as, specific customised equations for the different geographic regions which participated in their study. These equations provide flexibility for each ICU to select whether to use the overall global SAPS 3 equation or its own customised regional equation for the prediction of hospital mortality. The customised equations across geographic regions in SAPS 3 provide a local reference database that is useful for benchmarking purposes. On the other hand, the general equation of SAPS 3 provides a more generalised estimation and allows comparison with ICUs in other parts of the world since the global database included worldwide populations. The *logit* equation for global population is given as:

$$logit\,\mathrm{SAPS\,3\,global} = -32.6659 + ln\left(\mathrm{SAPS\,3\,score} + 20.5958\right) \times 7.3068, \qquad (2.5)$$

where the probability of hospital mortality can be computed using equation (2.3) (Moreno et al., 2005). The specific equations for different geographic regions are available in Appendix B (Table B11).

Moreno et al. (2005) found that the global SAPS 3 model exhibited good discrimination (with an area under ROC of 0.848) but poor calibration. They attributed this lack of calibration to be caused by an increased heterogeneity in sample due to inclusion of ICUs from all over the world. In addition, they explained that factors such as different genetic compositions, cultural and lifestyle differences, access to health care and differences in the use of major therapeutic measures indirectly contributed to the poor calibration. Differences in discrimination and calibration across different geographic regions were also observed in the study, although the majority of the seven regions recorded SMR values that were close to unity.

Soares and Salluh (2006) performed an external validation of the SAPS 3 Admission Score model specifically on cancer patients who were admitted to an oncologic ICU in Brazil. The model was found to be accurate and performed better than SAPS II in predicting hospital mortality in the Brazilian cohort of cancer patients. The SAPS 3 specific equation for Central, South America provided good calibration and discrimination, and was better than the global SAPS 3 equation. In a separate study in Belgium, Ledoux, Canivet, Preiser, Lefrancq and Damas (2008) concluded that SAPS 3 Admission Score model was superior to APACHE II, but not significantly better than SAPS II in terms of discrimination and calibration. The global equation of SAPS 3 did not fit their study data and overestimated hospital mortality. However, they found the calibration of SAPS 3 model customised for Central and Western Europe region to be adequate and more discriminative compared to the global SAPS 3 equation. As suggested by the above studies, customisation of regional equations may lead to improved accuracy in mortality prediction.

Metnitz et al. (2009) conducted a multicentre study involving validation and customisation of the SAPS 3 Admission Score model in a large cohort of intensive care patients in Austria. In the study, the global equation and customised Central and Western Europe equations had similar acceptable discriminative ability but inadequate calibration. Metnitz and colleagues claimed that the customised regional equations in SAPS 3 were not representative of all countries and may probably not be suitable for use in a particular country. In the event, where the customised regional equations failed to fit a particular country, they recommended development of country-specific equations. Using first level customisation approach, they proposed a customised equation specifically designed for Austrian patients. The relative weights of variables were maintained as in the original SAPS 3 Admission Score model, but the logistic coefficients were modified. The new customised model demonstrated excellent calibration on the general cohort of patients and various tested subgroups, as well as, maintained good discrimination as in the original model.

Lim et al. (2011) performed a retrospective study to validate the SAPS 3 global and Australasia regional predictive equations in a medical ICU in South Korea. Their findings revealed that although both equations produced good discrimination and calibration, the customised Australasia regional equation did not perform better than the global equation. One of the reasons cited for this finding was that the original Australasia regional equation was developed using a combined database of patients from Australia, Hong Kong and India. Differences in the genetic and cultural compositions between Australian and Asian patients were considered to have effect on the predictive accuracy of the Australasia regional equation.

The global predictive equation in SAPS 3 produced good discrimination and satisfactory calibration in two separate single-centre studies in Philippines (Hernandez & Palo, 2014) and India (Balaji, Rao, Kumar & Sammaiah, 2016). However, in both

studies, differences in the levels of calibration and discrimination across different subgroups of patients suggested that the overall performance of the global equation was affected by variations in case mix. In another recent study, Ma et al. (2017) evaluated the performance of the global SAPS 3 and Australasia regional equations in predicting hospital mortality in an emergency intensive care unit in China. Both the global and Australasia equations produced poor calibration and overestimated hospital mortality in the Chinese ICU. These studies suggested that the original SAPS 3 patient database was not representative of the global case mix and that there was a need to improve the accuracy of the regional equations.

2.2.4 Comparison of SAPS models

Although the three versions were uniquely different in their own ways, they shared some common objectives and features. First, as their names imply, all of the SAPS versions were designed to be simple and easy-to-use systems compared to other alternative scoring systems. This involved simplifying the process of data collection, in which all essential prognostic variables should be easily measured and detailed medical definitions should be made available to data collectors.

A comparison of physiological variables in SAPS models reveal a decreasing trend in the number of physiological variables used in the three versions. The original SAPS started with a total of 13 physiological variables, followed by 12 variables in SAPS II, and then only 10 variables in SAPS 3 Admission Score model. Spontaneous respiratory rate, haematocrit and serum glucose were only included in the original SAPS, but were not used at all in the subsequent revisions. Although variables such as urinary output, serum urea nitrogen, serum potassium, serum sodium and serum bicarbonate were used in both SAPS and SAPS II models, these variables were removed in SAPS 3 Admission Score model. The three new physiological variables that were

introduced in SAPS 3 Admission Score model were creatinine, hydrogen ion concentration and platelets.

All SAPS models considered age to have significant impact on mortality rate. Appendix B (Table B12) provides a direct comparison of the scores allocated to different age categories for the three SAPS versions. Chronic health variables were not part of the original SAPS model. However, three comorbidities (metastatic carcinoma, haematologic malignancy and AIDS) were introduced in SAPS II. The number of comorbidities was further expanded to five in SAPS 3 Admission Score model, with two more additional chronic disease being included (cancer therapy and cirrhosis). Besides the chronic health diseases, SAPS II and SAPS 3 Admission Score model required the type of admission to be recorded, i.e. whether the admission was scheduled surgical, medical or emergency surgery. This information was not captured in the original SAPS. Other new variables that were included in SAPS 3 Admission Score model included length of stay before ICU admission, intra-hospital location before ICU admission and use of major therapeutic options before ICU admission.

All equations and coefficients required to compute outcome probabilities for all the three SAPS versions are available in published journals. Substantial documentation and support for SAPS 3 Admission Score model is available at http://www.saps3.org. In conclusion, the performances of SAPS models have proven to be comparable to APACHE models. The most recent version, SAPS 3 Admission Score model, appears to show promising potential as a useful benchmarking tool for comparison of ICU performance. A summary of the differences among the three versions of SAPS is given in Appendix B (Table B13) for the reader's reference.

2.3 Mortality Probability Models (MPM)

2.3.1 MPM

Lemeshow et al. (1985) introduced Mortality Probability Models (MPM) as an alternative mortality prediction system to APACHE II. This model was developed based on 755 patients who were admitted to the Baystate Medical Center, US in 1983. The MPM models consisted of two unique models, namely the MPM_0-I admission model and MPM_{24}-I. The first model was developed to take into account information during ICU admission, while the second was used for prediction of mortality at 24 hours after ICU admission. At the time it was developed, the MPM_0-I was the first severity of illness scoring system that was derived at ICU admission. Its contemporaries, the APACHE, APACHE II, as well as, SAPS considered worst physiological measurements in the first 24 hours of ICU admission.

The MPM_0-I consisted of 7 independent admission variables that were recorded at the time of ICU admission or within 1 hour after ICU admission. These variables, which were non-dependent of treatment, were age, heart rate, level of consciousness, number of organ system failures, metastatic neoplasm, emergency admission and probable infection. The MPM_{24}-I also contained 7 variables that were measured at 24 hours after ICU admission. The variables were prothrombin time, urine output, creatinine, arterial oxygenation, continuing coma or deep stupor, confirmed infection and mechanical ventilation (Lemeshow et al., 1985).

Unlike the first generations of APACHE and SAPS, the MPM did not employ the use of clinical judgement for the selection and weighting of variables. Instead, linear discriminant function analysis with forward stepping was used to select significant predictors of hospital mortality. Age was considered to be a continuous variable while the remaining variables (physiologic, chronic diagnoses, acute diagnoses and surgical status) were classified as binary variables, where each variable was allocated a

coefficient and measurements were recorded as being present or absent. The dichotomous condition-based classification is clearly a unique feature of the MPM, since both the APACHE and SAPS systems measure the degree of physiologic derangement through the assignment of scores.

In contrast to the APACHE and SAPS systems, the MPM did not require computation of an intermediate severity of illness score for the calculation of mortality rate. The probability of hospital mortality can be calculated directly using a multiexponential equation based on the input data obtained from all the variables. MPM also did not require the selection of any primary diagnosis and utilised the fewest number of variables compared to APACHE, APACHE II and SAPS.

The calibration observed in MPM-I models were not satisfactory, thus these models were further refined and validated in 1988 (Lemeshow, Teres, Avrunin, & Gage, 1988). On top of the MPM_0-I admission model and MPM_{24}-I, a third model (MPM_{48}-I) was introduced to reflect mortality prediction at 48 hours after ICU admission. Several new variables were introduced in the revised admission model to account for respiratory failure, renal failure, emergency admission and surgical status, which were not included in the original MPM_0-I model developed in 1985. As for MPM_{24}-I model, the revision included additional variables such as shock during the first 24 hours after ICU admission and the number of lines observed at 24 hours. A variable to represent the use of continuous vasoactive drug therapy was included in the new MPM_{48}-I model. The complete list of variables used in the revised MPM_0-I, MPM_{24}-I, MPM_{48}-I models is given in Appendix C (Tables C1 - C3) respectively.

All these models require computation of a logit value that can be obtained through a general formula given as

$$logit = \hat{}_0 + \hat{}_1 x_1 + \hat{}_2 x_2 + ... + \hat{}_n x_n, \qquad (2.6)$$

where n = number of variables used in the model,

$\hat{\beta}_0$ = constant,

$\hat{\beta}_i$ = the estimated coefficient for the i-th variable (for $i = 1,...,n$)

$$x_i = \begin{cases} 0, & \text{absence of characteristic for all variables except age} \\ 1, & \text{presence of characteristic for all variables except age} \\ \text{age in years}, & \text{for age variable} \end{cases}.$$

The calculation for the probability of hospital mortality for the three MPM-I models is given as follows:

$$P(\text{hospital mortality}) = \frac{1}{1+e^{-logit}}. \qquad (2.7)$$

2.3.2 MPM-II

MPM-II was a part of Project IMPACT (Cerner Corporation, KS City, MO), which was initiated by the Society of Critical Care Medicine to address the need of a standardised data collection for mortality prediction. This updated version was published in 1993 and shared the same patient database that was used in SAPS II. This large international study involved 12,610 patients from two separate data sets. The first data set involved admissions in 6 ICUs in the US between 17 April 1989 and 31 July 1990. The second set involved admissions in 137 European and North American ICUs between 30 September 1991 and 10 May 1991. In order to avoid temporal bias due to the short interval of data collection, patients were randomly assigned numbers. Patients with values greater than 0.65 were placed in the validation sample, while the rest were included in the developmental sample. The study excluded patients who were under 18 years old, burn, coronary care and cardiac surgery patients. For patients with multiple ICU admissions, only the first admission was considered (Lemeshow et al., 1993).

The MPM-II models consisted of the admission model (MPM_0-II) and the 24-hour model (MPM_{24}-II). The number of variables in MPM_0-II admission model was kept to a minimum of 15 variables through application of multiple logistic regression

modelling. Compared to the 1988 MPM_0-I admission model, additional variables that were introduced in MPM_0-II admission model were cirrhosis, acute renal failure, cardiac dysrhythmia, cerebrovascular incident, gastrointestinal bleeding and intracranial mass effect. These variables were evaluated in the first hour before or after ICU admission. Each of these variables were recorded as absent or present and were assigned weights derived through logistic regression modelling. The calibration and discrimination of MPM_0-II in the validation sample was found to be good, with a *p*-value of 0.327 for the goodness-of-fit test and an area under ROC curve value of 0.824 (Lemeshow et al., 1993).

The MPM_{24}-II model was developed to provide updated probability of hospitality for patients who were still in the ICU after 24 hours of admission. A total of 13 variables were used in the model, with 5 being admission variables from MPM0-II and an additional 8 variables that were assessed at the 24-hour mark. The 5 admission variables were age, cirrhosis, intracranial mass effect, metastatic neoplasm and medical surgery admission. The additional variables that were recorded for the 24-hour model included coma/deep stupor at 24 hours, creatinine, confirmed infection, mechanical ventilation, partial pressure of oxygen, prothrombin time, urine output and continuous administration of vasoactive drugs. The MPM_{24}-II model displayed good discriminative ability and a high degree of calibration in the validation sample.

Lemeshow et al. (1994) conducted a study to evaluate whether the MPM_0-II and MPM_{24}-II models could be applied at other time periods, without further customisation. They concluded that models developed for use at specific time were not transferable to different periods than those in which they were developed. They further introduced two additional models, known as MPM_{48}-II and MPM_{72}-II, to provide prediction of mortality rates at 48 and 72 hours after ICU admission. These two models were exactly similar to the MPM_{24}-II in terms of variables and coefficients, except for the constant

terms. The constant terms were modified to reflect the positive relationship between probability of mortality and the length of stay in the ICU. The 48 and 72 hours models were useful to provide accurate mortality predictions that reflected changes in a patient's clinical profile according to response to therapy (Rue, Quintana, Alvarez, & Artigas, 2001). A comparison of variables and weightings in all of the four MPM-II models is provided in Appendix C (Table C4). The *logit* terms for all of the four models can be calculated using the approach given in equation (2.6). The probability of hospital mortality for these models is estimated using equation (2.3).

Castella, Artigas, Bion, and Kari (1995) performed an evaluation of MPM, MPM-II, APACHE II, APACHE III, SAPS and SAPS II models in a multicentre European and North American study. The study revealed that the performances of the newer models were much better than their older counterparts. APACHE II, SAPS, MPM_0-I and MPM_{24}-I exhibited poor calibration, despite showing acceptable discriminatory powers. Although the newer models (APACHE III, SAPS II and MPM-II) had comparable performances, none of them were found to be significantly better than the others.

Moreno, Miranda, Fidler, and Van Schilfgaarde (1998) performed a comparison between SAPS II and MPM_0-II admission models using the EURICUS-I (European Intensive Care Units Studies) database, which involved 16,060 consecutive admissions to 89 ICUs in 12 European countries. They concluded that both models presented poor calibration and overestimated risk of death in their patient database. The lack of calibration was attributed to either a shift in baseline characteristics or an improvement in the quality of care in Europe over the past five years, inability of models to adjust for case mix, and the presence of important clinical or nonclinical factors not accounted in the original models.

Using the same EURICUS-I database based on data collected from October 1994 to January 1995, Moreno and Apolone (1997) employed two different customisation strategies to the MPM_0-II admission model. The customisation strategies were used because MPM_0-II admission model did not perform well in the EURICUS-I population. The first strategy involved first-level customisation, while the second-level customisation was chosen as the latter approach. For the first-level customisation, the equation of the probability of hospital mortality was changed to

$$P_1(\text{hospital mortality}) = \frac{e^{-0.4926+0.7502(logit)}}{1+e^{-0.4926+0.7502(logit)}}, \tag{2.8}$$

where the *logit* term is given as in equation (2.6), with all the coefficients being the same as in the original MPM_0-II model. The second approach, however, was more complicated as it involved re-estimation of new coefficients for all the 15 variables defined in MPM_0-II model. A comparison of the original and new coefficients is given in Appendix C (Table C5). The new *logit* term for the customised model can be computed as

$$new\ logit = \hat{}_0' + \hat{}_1' x_1 + \hat{}_2' x_2 + \ldots + \hat{}_n' x_n, \tag{2.9}$$

where $n=15$, $\hat{}_0'$ and $\hat{}_i'(i=1,\ldots,n)$ are the new constant and new coefficients of the customised model. Moreno and Apolone (1997) found that second-level customisation was more effective and recommended this approach to address the problems of lack of fit in a given model.

On the other hand, a study by Arabi, Haddad, Goraj, Al-Shimemeri and Al-Malik (2002) indicated that MPM_0-II and MPM_{24}-II models performed well compared to APACHE II and SAPS II in a Saudi Arabian intensive care unit. Their findings revealed that MPM_0-II and APACHE II offered the best predictive accuracy, while MPM_{24}-II had the best calibration among the four models. In another single-centre study in Switzerland, Fischler et al. (2007) compared the performances of SAPS II,

MPM_0-II, MPM_{24}-II and Injury Severity Score (ISS) models in four well-defined patient subgroups (abdominal aortic aneurysm, multiple injuries, subarachnoid haemorrhage and head injury). MPM_{24}-II model demonstrated better predictive accuracy compared to SAPS II and Injury Severity Score (ISS). Fischler and colleagues believed that the superiority in performance shown by MPM_{24}-II was due to the model's ability to update mortality predictions based on changes in patient's conditions during the first 24 hours after ICU admission.

Glance, Osler and Dick (2002) reported that APACHE II, SAPS II and MPM_0-II substantially overestimated mortality in their data set, which consisted of ICUs which participated in the Project IMPACT group from 1995 to 1999. All of the three systems had poor calibration despite exhibiting good discrimination. They explained that the deterioration in performance of these three models was expected as they were developed using patient data sets that were more than ten years ago. In addition, they considered patient selection bias to be an influential factor since most of the ICUs that participated in the Project IMPACT might have chosen to report only the best-performing ICUs in the country. This eventually led to the development of the third generation of MPM model.

2.3.3 MPM_0-III Admission Model

Higgins et al. (2007) developed the MPM_0-III Admission Model in 2007 using a more contemporary Project IMPACT database, which consisted of 124,855 patients who were admitted to 135 ICUs at 98 hospitals. The study was conducted from October 2001 until March 2004 and involved mostly US hospitals, 3 Canadian hospitals and 1 Brazilian hospital. The developmental sample consisted of 60% of the total patients, while the rest were included in the validation sample. The study applied the same applicability criteria as the previous MPM_0-II version. Patients who were under 18 years of age, or suffering

from burns, acute myocardial infarction and cardiac surgery were excluded from the study.

MPM_0-III admission model consisted of 16 variables that were measured within 1 hour of ICU admission. All of the 15 independent variables used in the MPM_0-II were retained in this revision as they were found to be still relevant in predicting mortality risk. Only one additional variable was introduced in this new model, i.e. the code status of patients at the time of ICU admission, as this variable was found to have significant influence on mortality rate. The code status at the time of admission can be further classified into two categories ('zero factor' and 'full code'). As the study revealed that most of the elective surgical patients had no additional risk factors other than age, a 'zero factor' term was created to reflect a lower mortality risk for this subgroup of patients. The 'full code' term was used to indicate full resuscitation code status at admission and can be applied to patients with no therapy limitations (Higgins et al., 2007).

The MPM_0-III study also revealed the possibility of changes in the relative contribution of individual MPM_0-II variables over time. In the latest model, several variables such as gastrointestinal bleeding, chronic renal insufficiency, acute renal failure, cirrhosis and hypotension on presentation were assigned lower weights as they were considered to be less significant (Higgins, Teres, & Nathanson, 2008). The complete list of variables used in MPM_0-III and their corresponding weightings is available in Appendix C (Table C6). The model also included seven two-way interactions between age and coma/deep stupor, systolic blood pressure $\leq$ 90 mmHg, cirrhosis, metastatic neoplasm, cardiac dysrhythmia, intracranial mass effect and CPR prior to admission (see Appendix C (Table C7) for the coefficients of the interaction terms). These interaction terms were included as the MPM_0-III investigators found a negative relationship between age and the presence of comorbidity.

Higgins et al. (2007) observed that the previous MPM_0-II model overestimated hospital mortality when applied to a more contemporary Project IMPACT database. The new MPM_0-III admission model demonstrated acceptable calibration and good discrimination in the validation sample of this new database. To test the validity of the MPM_0-III admission model in an external population, Higgins et al. (2009) performed a separate validation study using a completely different data set from the one in which it was developed. The validation study included 103 ICUs which participated in Project IMPACT, with 24.3% being new ICUs that were not involved in the development of MPM_0-III admission model. The new database had higher admissions in patients with mechanical ventilation and lower admissions in patients with 'zero factors'. Despite these differences, MPM_0-III was proven to be robust and provided accurate predictions in the external validation set.

In a separate study using data from the CALICO (California Intensive Care Outcomes) project, Kuzniewics et al. (2008) evaluated variations in the mortality performances of APACHE IV, SAPS II and MPM_0-III models and concluded that there was substantial variation in ICU risk-adjusted mortality rates in the three models. APACHE IV was found to be the most superior model in terms of prediction accuracy and discrimination although all three models had comparable calibration. However, the time taken to collect data for APACHE IV was found to be twice as long for SAPS II and three times longer compared to MPM_0-III admission model. Kuzniewics and colleagues further recommended that APACHE IV be used only in cases with unlimited resources. The MPM_0-III admission model was proposed as a suitable alternative when restricted by cost and time constraints.

Riviello et al. (2016) conducted a validation of the MPM_0-III admission model in two public ICUs in Rwanda and found that the model did not discriminate or calibrate well in their cohort of study. Disparities in ICU settings, case mix, funding and

resources between a low-income country such as Rwanda and high-income countries such as the United States were considered important factors that contributed to the dismal performance of MPM_0-III admission model in the study. In a separate study, Lipshutz, Feiner, Grimes and Gropper (2016) compared the performance of the MPM_0-III admission model against the University Health Consortium expected probability of mortality (UHC EPM) model. Although the UHC EPM model was originally not designed for specific use in the ICU, the model performed better than the MPM_0-III admission model. The MPM_0-III admission model overestimated mortality in the study population, whereas the UHC EPM model did not overestimate mortality, making the UHC EPM model a better choice for benchmarking across ICUs.

Recently, Moralez et al. (2017) performed an external validation of SAPS 3 Admission Score Model and MPM_0-III admission model in 72 Brazilian ICUs. The large-scale study revealed that the SAPS 3 Admission Score model outperformed the MPM_0-III admission model, with better discriminatory power and calibration. The MPM_0-III admission model underestimated mortality, whereas SAPS 3 Admission Score model predicted outcomes accurately in the Brazilian ICUs.

2.3.4 Comparison of MPM models

In general, the development of MPM models took a completely different approach than that of APACHE and SAPS models. The MPM models were not based on a scoring system that assigned scores to clinical variables that have potential influence on predicted outcomes. Instead, development of MPM models involved a direct statistical modelling approach that eliminated the need of a scoring component.

The MPM models have proven to be useful in predicting mortality outcomes and are widely used within ICUs that participated in Project IMPACT. Changes in baseline characteristics of patients and improvements in quality of care have necessitated the

evolution of new generations of the MPM models. The uniqueness of MPM lies in its variants of models, namely the admission, 24-hour, 48-hour and 72-hour models. All of three generations of MPM systems included the admission model that reflected predicted outcomes that were independent of therapy. A comparison between the different generations of admission models revealed an increase in the number of variables being used (see Appendix C (Table C8)). The increase in number of variables with each model revision was necessary to achieve better model precision. A summary of the differences among existing MPM models is provided in Appendix C (Table C9).

MPM_{24}-I and MPM_{24}-II were useful in estimating probability of hospital mortality for patients who were still in the ICU after 24 hours. However, the 24-hour model was not updated in the third generation due to the availability of other robust models. Higgins et al. (2008) claimed that APACHE IV appeared to provide better and more accurate predictions within the first 24 hours interval compared to MPM_{24}-II. They also argued that modelling outcomes beyond 24 hours were impeded by factors such as sample size limitations, and the complex relationships between patient's response and treatment received. Although the latest MPM_0-III admission model performed well in the North American population, further assessment of its validity should be performed in external populations.

2.4 Comparison of APACHE, SAPS and MPM systems

The APACHE models are widely used in the US, since all of the four generations were developed and validated in the country. The first generation of SAPS was derived from a single-centre ICU in France, while the first generation of MPM was developed from a single-centre ICU in US. Patient characteristics in SAPS II and MPM II were quite similar since the two systems were built using the same European/North American Project IMPACT database. The latest MPM_0-III admission model was updated using a

more contemporary Project IMPACT database, while SAPS 3 Admission Score model was developed using worldwide populations.

Differences in patient selection contributed significantly to the accuracy of the predictive models. Thus, prior to conducting a performance comparison among several models, it is imperative to take note of the differences in the applicability criteria for each model. The first generations of APACHE and SAPS systems excluded paediatric and burns patients, whereas APACHE II excluded post coronary artery bypass graft patients. In the subsequent APACHE III and APACHE IV models, the exclusion criteria included those who were under 16 years old, with ICU length of stay not exceeding 4 hours, burns and acute myocardial infarction patients. Patients who went through coronary artery bypass graft were reported separately in APACHE III and APACHE IV. SAPS II, MPM-II and MPM_0-III admission models followed similar exclusion criteria and omitted patients who were under 18 years of age, or suffering from burns, coronary care or cardiac surgery. SAPS 3 Admission Score model, however, included coronary care and cardiac surgery patients, and excluded those who were under 16 years old.

APACHE and SAPS systems are similar in many ways, especially in the developmental process. Both systems consist of a scoring component, where an aggregate score is calculated to reflect the degree of severity of illness in critically ill patients. The main advantage of this approach is that patients can be stratified according to their scores, independent of the mortality prediction component. On the contrary, MPM was developed using a completely different approach compared to APACHE and SAPS. MPM is a condition-based system that requires assessment of variables in binary responses. Thus, MPM does not require the computation of an aggregate severity of illness score, as the probability of hospital mortality is directly estimated from input data.

Among the three systems, APACHE is the only one that requires specification of a unique primary diagnosis as reason for ICU admission. SAPS and SAPS II chose not to follow the same approach as APACHE due to problems associated with the identification of a single most important diagnosis. This is because incorrect assignment of diagnosis for patients who are suffering from multiple conditions may lead to inaccurate mortality prediction (Cowen & Kelley, 1994). However, reasons for ICU admission were included in SAPS 3 Admission Score model, but limited to only 10 diagnostic categories. This model also allowed simultaneous selection of several reasons for admission (Moreno et al., 2005). On the other hand, disease-specific information was deliberately left out in all of the MPM models so as to avoid inaccuracies that arise from misclassification of diagnosis (Higgins et al., 2007).

APACHE, SAPS and MPM are all physiological-based systems. Overall, APACHE and SAPS models have more common physiological variables compared to the MPM models. MPM has the least number of physiological variables among the three systems. A comparison of the number of physiological variables in the three systems is given in Appendix D (Tables D1-D3). The number of physiological variables and the need to specify ICU admission diagnosis contribute to additional burden in the data collection process for the APACHE models. In contrast, the data abstraction time for MPM models is the shortest since it involves data collection on only several physiological variables and entails only yes/no assessments.

The selection of variables and assignment of scores for the first generation of APACHE were done solely through clinical judgement. Although multiple logistic regression modelling was used to identify significant predictors in APACHE II and SAPS, assignment of scores was still based on clinical judgement for these two models. Variable selection and assignment of scores in APACHE III and APACHE IV were based on a combination of clinical judgement and statistical evaluation. Logistic

regression approach was used to guide variable selection and LOWESS method was applied to determine scores for the variables in SAPS II and SAPS 3 Admission Score models. As for MPM models, multivariable logistic regression was used to screen out non-significant predictors and to minimise the number of predictors in the models.

In practice, a good prognostic model should ideally have good calibration and discrimination that are reproducible and transportable (Justice et al., 1999). However, deterioration in performance over time is a common problem shared by all mortality prediction models. This is often indicated through overestimation of hospital mortality and poor uniformity of fit in a global population or certain subgroups. As discussed by other researchers in past validation studies, reasons for this trend can be due to changes in case mix of ICU patients or improvements in quality of care due to advances in medical technology and facilities over time. Models that were originally developed in a specific country may not be applicable in another country due to differences in patient characteristics, such as lifestyle and cultural differences, ethnic and genetic dispositions, as well as, the quality of medical services.

2.5 Reference Model

Despite differences in terms of variables and modelling approach, the latest generation of ICU prognostic models appear to demonstrate reliable and convincing results in their original developmental samples. Previous studies have suggested that older models that were developed a long time ago may no longer be accurate when applied on a recent database. The latest generation of models such as APACHE IV, SAPS 3 Admission Score Model and MPM_0-III admission model are expected to provide better estimates since they were developed using larger multi-centre databases and advanced statistical methods. Based on this argument, the best option was to choose one of these latest models to be the reference in this study. A simultaneous implementation and

comparison of the three latest models was not possible in this study due to challenges in the data acquisition process, which involved differences in variables and interpretation of medical terms among the three models.

Upon deliberation of the advantages and limitations of the three latest models, APACHE IV was chosen as the reference model in this study. Selection of this model was principally driven by the fact that APACHE IV had the highest number of physiological variables among its contemporaries. The availability of more information on the physiological conditions of the ICU patients was deemed beneficial for the model development phase in this research. In addition, APACHE IV required data collection throughout the first day after ICU admission, whereas SAPS 3 Admission Score and MPM_0-III admission models relied on data that were collected within the first hour before/after ICU admission. The complete availability of patient information was a concern in SAPS 3 Admission Score and MPM_0-III admission models due to the shorter interval of data collection. The data collection interval for APACHE was more reasonable as it allowed all important physiological attributes to be documented.

The decision to use APACHE IV as the reference model was also due to its promising potential shown in several recent external validation studies. Positive findings by Kherallah et al. (2008) and Bhattacharyya et al. (2009) suggested the possibility of APACHE IV being robust enough for application in other countries and in single-centre settings.

PART TWO: LITERATURE REVIEW ON STATISTICAL MODELLING

2.6 Modelling of ICU risk of death using logistic regression approach

Regression models are extensively employed as forecasting or predictive tools due to their ability to describe the relationship between a response (dependent) variable and a set of explanatory (independent) variables. Traditionally, linear regression models are applied to cases where the outcome variable is continuous, with the variability of outcome being assumed as constant for all values of predictor variables. Nevertheless, linear regression models may not be appropriate when the response variable is categorical or specifically dichotomous in nature (Czepiel, 2002). In such circumstances, logistic regression is preferable because it allows the use of continuous or categorical predictors and provides the ability to adjust for multiple predictors (LaValley, 2008). In addition, the logistic regression model also allows the use of design (dummy) variables to represent nominal scale variables.

As such, logistic regression models are suitable to be used as predictive models that involve binary outcomes, especially in medical and epidemiologic research. For instance, the outcome variable can represent the occurrence of an event of interest, such as the presence or absence of a disease or death (Anderson, Jin, & Grunkemeier, 2003). Modelling techniques that are based on logistic regression approach have been employed in the development of the latest versions of ICU prognostic models. These models applied logistic regression approach in the selection and weighting of individual predictors.

A binary logistic regression model is used to analyse the relationship between a dependent variable (probability of a dichotomous outcome) and multiple independent variables (risk predictors) through a logistic functional form. In a logistic regression model, a natural logarithm transformation is applied to the odds of the outcome

variable, where the odds of an outcome occurring are defined as the ratio of probability of outcome occurring to probability of outcome not occurring (Peng, Lee, & Ingersoll, 2002). This results in a *logit* term that is a linear function of the independent explanatory variables. The *logit* term for a multiple logistic regression model can be expressed as

$$f(\boldsymbol{x}) = \ _0 + \ _1 x_1 + \ _2 x_2 + ... + \ _p x_p = \sum_{j=0}^{p} \ _j x_j, \tag{2.10}$$

where $\boldsymbol{x}' = (x_0, x_1, x_2, ..., x_p)$ is the vector of independent variables and $\ _j$, for $j = 0,1,...,p$ represents the unknown parameters to be estimated by the model. In ICU prognostic models, the covariates $\boldsymbol{x}' = (x_0, x_1, x_2, ..., x_p)$ refer to the mortality risk predictors, while $\ _j, j = 0,1,...,p$ represent their corresponding regression coefficients, with $\ _0$ being the intercept value. The predicted probability of mortality is evaluated as

$$®(\boldsymbol{x}) = \frac{e^{f(x)}}{1 + e^{f(x)}}. \tag{2.11}$$

In a logistic regression model, the conditional mean of the dichotomous outcome variable is assumed to follow a Binomial distribution (Peng et al., 2002). Unlike ordinary least squares regression, logistic regression does not require any assumptions to be made about the distribution of the explanatory variables, i.e. in terms of normality of error distribution, homoscedasticity of errors and linearity between dependent variable and independent variables (Park, 2013). However, logistic regression requires observations to be independent, absence of multicollinearity among the independent variables, and linearity assumption in the *logit* for continuous variables (Stoltzfus, 2011). Logistic regression requires large sample sizes in order to provide adequate statistical power (Park, 2013). In addition, Peduzzi, Concato, Kemper, Holford and Feinstein (1996) advocated the use of 10 events per variable as a rule of thumb in a logistic regression model.

2.6.1 Parameter Estimation in Logistic Regression using Maximum Likelihood Estimation (MLE) approach

Estimation of the coefficients in a multivariable logistic regression is often complex and requires the use of iterative techniques, as it involves equations that cannot be solved explicitly (Bagley, White, & Golomb, 2001). Maximum likelihood method is usually employed as a standard approach in parameter estimation in a logistic regression model. Derivation of the maximum likelihood equation comes from the probability distribution of the outcome variable. The frequentist maximum likelihood method assumes that available data are random subsets from a population of interest, where the unknown parameters are assumed to be fixed with an associated random error (Hamra, MacLehose, & Richardson, 2013). For a given data set and underlying probability model, the maximum likelihood method finds estimates that maximise the probability in obtaining the data set (Cole, Chu, & Greenland, 2014). The maximum likelihood method generates point estimates and their corresponding standard errors.

Maximum likelihood estimation (MLE) approach is available in most statistical software packages and is commonly used in parameter estimation in ICU prognostic models. The MLE approach has desirable properties such as consistency, asymptotic normality and asymptotic unbiasedness when the sample size is sufficiently large, where the estimated coefficients asymptotically approach the population values as sample size increases (Cole et al., 2014). Due to its asymptotic behaviour, the maximum likelihood method produces poor and unreliable results in terms of *p*-values and parameter estimates for small sample sizes (King & Ryan, 2002). The size of bias of a maximum likelihood estimator is affected by sample size, where the odds ratios in a logistic regression model are usually overestimated in small to moderate sample size studies (Nemes, Jonasson, Genell, & Steineck, 2009). Long (1997) recommended using maximum likelihood estimates for sample sizes above 500, and not for studies with less

than 100 samples. The MLE approach is also not recommended for sparse or unbalanced data sets, and for cases where the majority of outcome values are either 0 or 1 (Mehta & Patel, 1995).

2.6.2 Parameter Estimation in Logistic Regression using Bayesian Markov Chain Monte Carlo (MCMC) approach

Estimation and inference of model parameters can also be alternatively done using Bayesian approach. The Bayesian approach in model estimation involves the process of fitting a probability model for a given data set through the estimation of model parameters via sampling (Congdon, 2001). Unlike classical frequentist approach, Bayesian approach treats the unknown parameters in a model as random variables and uncertainty is measured using the posterior distribution of a parameter (Ntzoufras, 2009). Statistical inference on model parameters is based on the posterior distribution, which provides information about the parameters in a statistical model, given the observed data and existing knowledge. Bayesian approach takes into account knowledge from past experience or studies and current sample data from a population of interest to make statistical inference, and uses information provided by observed data to update estimates of the unknown parameters in a model.

Computation of the posterior distribution involves integral solutions to marginal distributions. For high dimensional problems, exact inference on the posterior distribution is not possible since the integrations are intractable through analytical methods. Stochastic simulation-based methods such as the Markov Chain Monte Carlo (MCMC) algorithms (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) are commonly applied in Bayesian analysis for estimation of posterior distributions, where analytical solutions are difficult to obtain. The Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) and Gibbs Sampling algorithm

(Gelfand & Smith, 1990; Geman & Geman, 1984) are two widely-used MCMC algorithms.

MCMC is a stochastic method used to simulate random samples that would form an irreducible Markov chain, whose stationary distribution is exactly the same as the posterior distribution (Geyer, 1992). The Markov chain process simulates samples sequentially from a posterior distribution, beginning with an initial value that is sampled from the initial distribution. Successive samples are generated by using previous sample values, where the only information required for the prediction of a future value is the current state of the random variable, whereas knowledge of past states is not required. Iteration of this simulation procedure is continued for a long time until the ergodic chain converges to its stationary distribution. Monte Carlo integration is then applied to generate summary statistics from the random samples (Hamra et al., 2013). Summary measures of the posterior distributions include the posterior means and the 95% credible intervals, which are used as alternative measures to the maximum likelihood estimates and 95% confidence intervals (Dunson, 2001).

There are several advantages in employing a Bayesian MCMC approach in model estimation and inference. First, the flexibility of the Bayesian approach to update prior information about the underlying parameters with information from cumulative or past experience is considered one of its advantages. This allows information from prior existing models to inform a more current model. The Bayesian approach also provides a degree of uncertainty in the model, thus yielding predictions that are more realistic and safeguards against overfitting of models more than frequentist approaches. In addition, the Bayesian approach relies on exact inference, instead of large sample asymptotic approximations. This facilitates an easier and more intuitive interpretation of the credible intervals of the estimated parameters of a predictive model.

2.7 Bayesian Markov Chain Monte Carlo approach in prognostic modelling

The Bayesian MCMC approach can be applied in large classes of problems such as in hierarchical models, generalised linear models and time series models. These methods are suitable for applications in medical and biostatistics research, financial econometrics, actuarial science, spatial statistics and genetics. Difficulties in performing Bayesian analyses for multivariate data and the availability of other easier techniques are some of the reasons why the Bayesian approach is less popular compared to other traditional methods. Application of this approach also requires the use of specialised software and the knowledge to perform MCMC analyses. Although this approach is widely used for predictive analysis in medical applications, its use in predicting in-hospital mortality has been rather underutilised.

A search through the current literature revealed only a handful of studies that involved the use of Bayesian MCMC methods for prediction of in-hospital risk of death in specific diseases and subgroups of patients. These studies were mostly focused on the application of Bayesian MCMC methods in variable selection and model choice. In addition, these studies also discussed the problems in eliciting prior distributions for regression parameters in order to obtain posterior distributions.

Chen, Ibrahim and Yiannoutsos (1999) offered some discussion on the difficulties in specification of prior distributions for regression parameters in the logistic regression model, and issues related to variable selection. They proposed a family of informative prior distributions for the logistic regression model and designed a methodology for use in cancer and AIDS clinical trials research. Computation of the posterior probabilities was performed based on a single Gibbs sample from the full model.

In another study, Bedrick et al. (1997) presented a simple approach to a fully Bayesian analysis of a binomial regression model to predict survival rate of trauma

patients. They followed the approach by Tsutakawa and Lin (1986), by eliciting information about success probabilities $\tilde{p}_i$ at selected covariate vectors $\tilde{x}_i$, and using these at a later stage to include the essential prior on regression coefficients $_i$ for the binomial regression model. A variant of importance sampling method was used to obtain inferences on the posterior distribution.

Souza and Migon (2004) developed a Bayesian binary regression model to predict in-hospital mortality of patients after acute myocardial infarction (AMI) using MCMC methods for inference. The approach that was used in the model building process was largely based on a modified form of the Hosmer and Lemeshow (2000) approach. Souza and Migon (2004) proposed the use of Bayes factor in the selection of different competing multivariable models. Bayes factor is a ratio of marginal likelihoods that is used in comparing two models, i.e. Model M_1 against Model M_0. The formula to compute Bayes factor is given as

$$B_{1,0} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_0)}, \tag{2.12}$$

where $p(\boldsymbol{y}|M_j)$ is the marginal probability of obtaining data $\boldsymbol{y}$ from model M_j, for $j = 0,1$. The best model was selected based on a simple rule of thumb that was proposed by Kass and Raftery (1995). Souza and Migon (2004) adopted the Gibbs sampler approach using BUGS software to obtain information regarding the posterior distribution of interest and performed residual analysis to evaluate for model accuracy.

In another study, Kazembe, Chirwa, Simbeye and Namangale (2008) employed fully Bayesian MCMC simulation techniques in developing semiparametric regression models for analysis of risk of malaria-related hospital mortality. They selected diffuse (non-informative) prior distributions for the regression parameters, which assumed the prior density of the parameters to be constant. This approach was chosen due to lack of specific information regarding the regression parameters. Analysis of the posterior

distribution of model parameters was performed using a hybrid MCMC sampling scheme of iteratively weighted least squares and the Metropolis-Hastings algorithm.

In a recent study in Malaysia, Chiaka, Adam, Krishnarajah, Shohaimi, and Guure (2015) compared the performance of two Bayesian logistic regression models with the frequentist logistic regression model in predicting risk factors of Type 2 diabetes mellitus. The first Bayesian model involved the use of a non-informative flat prior distribution, whereas a non-informative not perfectly flat prior was employed in the second Bayesian model. Their results indicated that the first Bayesian model with non-informative flat prior produced results that were comparable to the frequentist MLE logistic regression model. Improvement in results was observed through the use of a non-informative not perfectly flat prior in the second Bayesian model.

Bayesian networks approach is an alternative to logistic regression model and is also popularly used for mortality and disease prediction. Naive Bayesian classifiers can be used in variable selection and development of mortality prediction models. For instance, Ryynänen, Soini, Lindqvist, Kilpeläinen, and Laitinen (2013) developed a Bayesian mortality prediction model for patients with chronic obstructive pulmonary disease (COPD). The study used a naive Bayesian classification approach to determine the risk factors of mortality and very poor patient's health related quality of life in the cohort of COPD patients. Naive Bayesian classification approach have also been applied in the development of risk prediction models for patients with specific diseases such as lung and breast cancer (Martin-Sanchez et al., 2016), coronary heart disease (Wang et al., 2016), cirrhotic (Blanco, Inza, Merino, Quiroga, & Larranaga, 2005).

From these studies, it is evident that Bayesian methods can be utilised in different ways in predicting hospital mortality, and provide an alternative approach in model estimation. However, the Bayesian approach is often seen as being complicated, especially for models that involve many variables, as in the case of ICU prognostic

models. In this study, the Bayesian MCMC approach was chosen to develop logistic regression models for prediction of mortality risk in an ICU, particularly in estimation and inference of model parameters.

There are plenty of issues and considerations that need to be addressed in the modelling process that involves a Bayesian MCMC approach. Specification of prior distributions for the regression parameters is one of the biggest challenges in model development, especially if the model has a large number of variables. Prior distributions for model parameters are specified through the prior mean and variance. Non-informative priors are usually used for cases where there is little information on the regression parameters. Diffuse priors are non-informative priors that assume the prior density of the parameters to be constant. They are often used in studies that involve large samples and where the posterior on regression coefficients is approximately normal (Bedrick et al., 1997). The accuracy and efficiency of estimation in a model can be improved through the use of informative priors, which incorporates information from past studies into the current study (Dunson, 2001).

Other aspects to be considered in Bayesian MCMC modelling include determining the stopping criteria in terms of how long a Markov chain simulation should be run, as well as, monitoring of chain convergence. Ideally, the number of iterations for an MCMC routine should be as large as possible to achieve model convergence. Advances in computing power have made this possible, where it is possible for the number of iterations to reach up to millions of runs for models that are not that complicated.

Computing efficiency can be enhanced by choosing initial values of the Markov chain that are near to the centre of the target posterior distribution. This is to avoid problems such as slow convergence and excessive iterations in certain parts of simulation (Gelman, 1996). Besides that, samples that are obtained during the initial

burn-in period of the chain should also be discarded from analysis because these values may affect accuracy in estimation of posterior distributions (Ntzoufras, 2009).

Many researchers have debated on the use of a single chain against several parallel chains in Bayesian MCMC simulations. Geweke (1992) advocated the use of a very long run on a single chain since this method is able to find new posterior modes. A single long chain is also supposed to be more efficient due to the requirement of only one transient phase (Smith & Roberts, 1993). On the other hand, Gelfand and Smith (1990) recommended the use of many short chains, while Gelman and Rubin (1992) proposed using several long runs in a small number of independent chains. Multiple chains allow comparison in the convergence of the chains, where convergence is monitored through analysis of variance between and within the chains.

The availability of software and advances in computing technology help to improve the efficiency of Bayesian analysis using MCMC techniques and provide solutions to highly complicated problems. Nevertheless, users of these programs have to exercise caution in using and interpreting the results of analysis, in addition to being aware of the underlying assumptions involved in the development of the software.

2.8 Mortality Prediction using Decision Tree approach

A decision tree is commonly used as a classifier in the fields of machine learning and data mining (Badriyah, Briggs, & Prytherch, 2012; Meyfroidt, Güiza, Ramon, & Bruynooghe, 2009). A decision tree is graphically illustrated as hierarchical nodes that form a tree diagram. The single node at the top of the tree is known as the root node. Below the root node are other child nodes that represent attributes, events or possible outcomes of decision rules. These nodes are connected to each other through incoming or outgoing edges. A node that has outgoing edges to other nodes is known as an internal node. The internal node splits the input data set into mutually exclusive and

collectively exhaustive segments according to specified decision rules. The endpoints of the decision tree that have no more child nodes are defined as the terminal nodes or leaves. The path from each terminal node to the root node is unique. In the case of a mortality prediction model, the probability of death for a patient with specific conditions is given in the terminal node or leaf. Induction of the decision tree is based on a training set, whereas validation of the decision tree model is done using a different set of data (Podgorelec, Kokol, Stiglic, & Rozman, 2002).

CART (Classification and Regression Trees) (Breiman, Friedman, Stone, & Olshen, 1984) and CHAID (Chi-Squared Automatic Interaction Detection) (Kass, 1980) are two widely-used classification algorithms used in construction of decision trees. These classification algorithms are available in most statistical software packages. CART algorithm is able to handle categorical or continuous variables and can be used for regression and classification (Loh, 2011). The approach is based on binary recursive partitioning, where each internal node can be split into two child nodes. Splitting of nodes in CART is based on a Gini index and covers an exhaustive search of all possibilities. The tree building process is stopped when a maximal tree is produced. Cost-complexity pruning is employed to reduce the size of decision tree, as a large tree is associated with overfitting of model (Rokach & Maimon, 2008).

On the other hand, CHAID algorithm can be used to construct non-binary trees, and uses multiway splits at each node (Kim & Loh, 2001). This method is suitable for nominal, ordinal and continuous data, where continuous predictors are split into categories with approximately equal number of observations (Nisbet, Elder, & Miner, 2009). CHAID relies on a chi-square test to detect interactions between variables and to determine the next best split at each node. The advantage of CHAID is that at each step, the algorithm selects independent predictors that have the strongest association with the dependent variable (Badriyah et al., 2012). Moreover, the method does not require post-

pruning, and prevents against overfitting by not producing trees that are too large (Rokach & Maimon, 2008). Relationships between variables are easily visualised and interpreted from a decision tree generated using CHAID algorithm. In addition, it is able to handle cases of missing data (Song & Lu, 2015).

In the medical community, decision trees provide an alternative to logistic regression in the modelling of probabilistic clinical outcomes (Badriyah et al., 2012). The decision tree induction approach can be used to develop clinical decision rules and to predict continuous or categorical outcomes. Thus, this method is suitable for predicting mortality risk in ICUs. Decision tree induction approach has been widely applied to predict mortality risk in various groups of patients. Some of these studies involved patients with cancer (Hess, Abbruzzese, Lenzi, Raber, & Abbruzzese, 1999; Mohammadzadeh, Noorkojuri, Pourhoseingholi, Saadat, & Baghestani, 2014), carpal tunnel syndrome (Rudolfer, Paliouras, & Peers, 1999), heart failure (Ebell, 2007; Zhang, Goode, Rigby, Balk, & Cleland, 2013) and septic shock (Wong et al., 2014).

Over the years, several studies have reported positive outcomes in the use of decision tree models in predicting intensive care unit mortality risk. For instance, de Rooij, Abu-Hanna, Levi and de Jonge (2007) developed a classification tree to predict mortality risk of a cohort of 6,867 elderly intensive care unit patients from 21 Dutch ICUs. The performance of the classification tree was compared against SAPS II model, where the former approach was found to be slightly superior in terms of positive predictive values and prediction accuracy for patients with very high risk. In another study involving 38,474 admissions to the University Kentucky Hospital in USA, Kim et al. (2011) concluded that the decision tree model outperformed the conventional logistic regression model of APACHE III in terms of discrimination.

Many studies have debated on the differences between the conventional logistic regression approach and the decision tree method (Badriyah et al., 2012; Kim et al.,

2011; Long, Griffith, Selker, & D'Agostino, 1993; Perlich, Provost, & Simonoff, 2003). There are advantages and disadvantages in both methods and at best, they can be used to complement each other. The non-parametric nature in the decision tree approach is more appealing because it does not require linearity assumptions to be met. Ease of interpretation and its robustness to noise and incorrect classification are other advantages of the decision tree approach (Meyfroidt et al., 2009). The decision tree method is also able to handle cases of missing data in two ways. The first way is to classify the missing values in a separate node that can be analysed with other nodes. Another way is to construct a decision tree model that assigns the variable with many missing values as a target variable for prediction, and to then use the predicted values to replace the missing values (Song & Lu, 2015).

A disadvantage of decision tree is that is it prone to overfitting (Abu-Hanna & de Keizer, 2003). Other criticisms of the decision tree approach include the requirement for segments to be mutually exclusive, issues of missing data, complexity of decision tree and the stopping criteria (Kim et al., 2011). Perlich et al. (2003) claimed that the decision tree approach was more appropriate for studies with large sample size, whereas the logistic regression approach was more suitable for smaller training data sets.

2.9 Assessment of Model Accuracy

The accuracy of prognostic models is commonly assessed through measures of discrimination and calibration. Discrimination refers to the ability of the model to classify and segregate patients who will survive from those who will die. Calibration is used to assess the agreement between predicted mortality risk obtained from a model and the observed mortality risk.

2.9.1 Model Discrimination

The assessment of discrimination is usually performed using the receiver operating characteristic (ROC) curve, which is widely used in radiology diagnostic applications to distinguish between patients who are diseased and non-diseased (Goodenough, Rossmann, & Lusted, 1974). A receiver operating characteristic curve is a graphical plot of test sensitivity against its false positive rate for all possible cut points, where false positive rate is defined as (1 – specificity). The accuracy of a diagnostic test is defined by sensitivity and specificity measures. Sensitivity refers to the number of true positive decisions, whereas specificity is the number of true negative decisions (Obuchowski, 2003).

The area under a receiver operating characteristic curve (AUC) is a combined measure of sensitivity and specificity, and can be interpreted as a probability of correct classification or prediction (Hanley & McNeil, 1982). This measure can be used in prognostic models to assess how well a model is able to discriminate between hospital deaths and survivors (Grunkemeier & Jin, 2001). The range of AUC value is between 0.0 and 1.0. In theory, an AUC value of 1.0 indicates perfect discrimination and implies that the predicted outcomes will be exactly the same as the observed outcomes for all patients. In contrast, an AUC value of 0.0 suggests the absence of any discriminatory power, resulting in totally incorrect predictions for all patients. Higher values of AUC indicate better discrimination. The practical lower limit for AUC is 0.5, and is represented by a diagonal line on the ROC curve. AUC value of 0.5 implies that the model's discrimination between deaths and survivors is no better than chance.

2.9.2 Model Calibration

The Hosmer-Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1980) is a popular measure of calibration used in ICU mortality prediction models. This method evaluates

how well a model is calibrated by measuring the degree of agreement between observed and predicted mortality probabilities across different subgroups of patients in a model. A model is considered to be well-calibrated if the model predicts the numbers of observed and expected mortality equally well across different groups (Randolph, Guyatt, Calvin, Doig, & Richardson, 1998).

The Hosmer-Lemeshow goodness-of-fit test requires construction of *n* subgroups for different categories of patients, where the predicted and observed mortality probabilities will be compared within each subgroup (Hosmer & Lemeshow, 1980). The Hosmer-Lemeshow goodness-of-fit test statistic is computed as

$$\hat{C} = \sum_{k=1}^{n} \frac{(o_k - e_k)^2}{e_k\left(1 - \frac{e_k}{n_k}\right)}, \tag{2.13}$$

where *n* is the number of groups, o_k is the observed number of outcomes in the *k*-th group and e_k is the expected number of outcomes in the *k*-th group. The values of o_k and e_k are defined as $o_k = \sum_{j=1}^{n_k} y_{kj}$ and $e_k = \sum_{j=1}^{n_k} \hat{\pi}_{kj}$ respectively, where y_{kj} denotes the outcome probability for observation j in group k, $\hat{\pi}_{kj}$ is the estimated probability for observation j in group k, and n_k is the sample size for group k. The $\hat{C}$ statistic is known to follow a chi-squared distribution with (*n* - 2) degrees of freedom when the fitted logistic regression model is accurate (Hosmer & Lemeshow, 1980). A model is considered to have good overall fit if the *p*-value for the Hosmer-Lemeshow test exceeds 5% level of significance in this study. In addition to the Hosmer-Lemeshow test, calibration curves can also be used to compare observed and predicted risk of deaths across different subgroups of patients (Steyerberg et al., 2010). The calibration curve is a graphical plot of the mean predicted probabilities against the mean observed outcomes for the different subgroups of patients (Cook, 2006).

Other measures that can be used to evaluate overall model fit are the Standardised Mortality Ratio (SMR) and Brier score (Brier, 1950). The Standardised Mortality Ratio is defined as the ratio of mean observed deaths in a study population over the mean predicted deaths in the same population (Goldman & Brender, 2000). A value of 1.0 indicates a perfect fit, where the predicted and observed death rates are the same. A model is considered to overestimate mortality risk when the SMR value is less than 1.0. Conversely, the model underestimates mortality risk in a population of study when the SMR value is greater than 1.0.

SMR is commonly used as a benchmarking index to evaluate and compare efficiency and performance of ICUs (Zimmerman, Alzola, & Von Rueden, 2003). SMR can also be used to compare the relative performance between different ICUs in the one country or across different countries. This index can also be used to compare the performance of an ICU over time, and provide a mechanism to reflect the prediction accuracy of the model over a long period of time.

Brier score is a common scoring rule used for weather probability forecast in meteorology (Brier, 1950). In prognostic models, Brier score can be used to measure the squared distance between observed and predicted probabilities for each patient. The Brier Score for each model was calculated as

$$\text{Brier score} = \frac{1}{n}\sum(\hat{y}_i - y_i)^2, \quad i = 1,...,n, \tag{2.14}$$

where $\hat{y}_i$ is the predicted probability of mortality estimated by the model, y_i is the actual outcome for patient i, and n is the number of patients. The decision space for a useful model was restricted to (0, 0.25). In the worst-case scenario, a non-informative model with each predicted probability being set as 0.50 will produce a Brier score of 0.25. A model with a smaller Brier score was considered to have better accuracy (Gerds, Cai, & Schumacher, 2008).

Overall, model accuracy is assessed through discrimination and calibration properties. Prior to application of a model in a certain population, the model should demonstrate sufficient discrimination and calibration since model accuracy will be compromised if these two criteria are not satisfied.

CHAPTER 3: METHODOLOGY

This chapter covers the design, methodology and direction of this research. The first part explains the data collection, methodologies and framework employed in the reference model. A detailed discussion on the statistical methodologies employed in the modelling phase of this study is also presented in this chapter. For ease of understanding, Table 3.1 provides a summary of the research objectives and the corresponding methodologies employed to achieve the objectives.

Table 3.1: Summary of research objectives and the methodologies employed.

Research Objective	Methodology
To identify and choose a suitable recent ICU prognostic model to be used for reference in a particular Malaysian ICU by performing a comprehensive review of existing well-established ICU prognostic models.	▪ Literature review on existing ICU prognostic models.
To investigate the validity and accuracy of the chosen model in a Malaysian ICU by performing an external validation of the chosen reference model.	▪ Computation of APS. ▪ External validation of APACHE IV in a Malaysian ICU. First-level customisation strategy on APACHE IV model.
To determine the limitations and gaps in the statistical methodology of the reference model, and identify areas for improvement.	▪ Literature review on APACHE IV. ▪ Literature review on alternative modelling strategies for model development.
To propose alternative techniques in the modelling of ICU mortality risk.	▪ Development of five Bayesian logistic regression models. - Hosmer-Lemeshow (2000) modelling strategy. - variable selection based on likelihood ratio test and Bayesian credible interval. - model selection based on Deviance Information Criterion (DIC). ▪ Development of a decision tree model using CHAID algorithm.
To compare the performance of models developed using alternative modelling strategies against a model developed using a frequentist approach.	▪ Compare performance of Bayesian model against frequentist MLE logistic regression model.
To propose the best model for prediction of individual mortality risk in a Malaysian ICU.	▪ Assess overall performance measure in terms of discrimination and calibration.

3.1 Design and Setting

An independent prospective observational study was conducted in the Hospital Sultanah Aminah Johor Bahru (HSA) ICU between the period of 1 January 2009 and 30 June 2010. HSA is a government tertiary referral hospital that was established in the 1940s in Johor Bahru, Malaysia. The hospital has a single multidisciplinary intensive care unit that is considered one of the largest in Johor Bahru, with a current bed size of sixteen. All of the beds are equipped with mechanical ventilation facility. The HSA ICU provides services to general medical, surgical and trauma patients. This ICU does not admit post-coronary artery bypass graft (CABG) patients because they are treated in a separate unit in the hospital.

3.1.1 Patient selection and exclusion criteria

The subjects of this study were defined as critically ill adult patients who were admitted to the single multidisciplinary ICU in HSA between 1 January 2009 and 30 June 2010. Following the exclusion criteria in APACHE IV, these groups were excluded:

i) patients with age less than 16 years old,

ii) burn patients,

iii) patients with less than 4 hours of ICU stay,

iv) transplant patients,

v) patients with more than 365 days of hospital stay

vi) patients with missing day 1 APS, and

vii) transfer cases from another ICU.

For patients with records of multiple admissions, only the first ICU admission was taken into consideration. Post-coronary artery bypass graft patients were referred to a separate unit and were in the exclusion list.

3.1.2 Data collection and variables

This research was approved by the Medical Research and Ethics Committee, Ministry of Health, Malaysia. The requirement for informed consent from all participants was waived because data collection was based on existing medical records and did not involve any clinical intervention. All patient records and information were de-identified and analysed anonymously. Data collection for the entire study was physically performed by HSA ICU nurses. These data were then manually transferred to an online-computerised database by the hospital's medical officers. To ensure data integrity and traceability, separate user accounts and passwords were issued to each data entry personnel.

Data collection was based on the APACHE IV approach. However, frequency of data collection followed the current practice in HSA ICU. The following items were collected for the study and are summarised in Table 3.2:

i) demographic details (patient's age, gender and ethnic group),

ii) admission (time and date of admission, source, primary ICU admission diagnosis),

iii) chronic health conditions (all variables defined in APACHE IV and diabetes)

iv) physiological and laboratory measurements (all variables in APACHE IV)

v) ICU length of stay, and

vi) discharge data (vital outcome status, discharge location).

Patient's details such as demographic, admission data, presence of comorbidities, operative status, physiological and laboratory measurements were taken at the time of ICU admission. Information on whether the patient was suffering from diabetes mellitus was also collected, although this variable was not included in APACHE IV. Throughout the course of patient's stay in the ICU, physiological measurements (heart rate, systolic blood pressure, mean blood pressure, diastolic blood

pressure, temperature, mechanical ventilation status, total respiratory rate, Glasgow Coma Scale (GCS) score, glucose and urine output) were monitored on an hourly basis. On the other hand, laboratory measurements (haematocrit, white blood cell count (WBC), blood urea nitrogen (BUN), creatinine, sodium, bilirubin and albumin) were monitored less frequently, approximately twice in a day. Assessment of patient's neurological functions (arterial-ph) and blood gases (PaO_2 and FiO_2) was performed between four to six times daily. The admission diagnoses for each patient were determined by the ICU specialist on duty and subsequently verified by an intensivist.

Table 3.2: Data items collected within first day of admission in HSA ICU.

Variable	Description
Age	Continuous measure, in years
APS	Continuous measure, sum of scores for the worst values of each physiological variable within the first day of ICU admission
Pre ICU length of stay	Continuous measure, in days (square root)
PaO_2:FiO_2 ratio	Continuous measure
Gender	Categorical: male (reference), female
Ethnic group	Categorical: Malay (reference), Chinese, Indian, Others
ICU admission source	Categorical: Floor/ward (reference), Another special care unit, Operating room
Chronic health	Categorical: None (reference), AIDS, cancer, cirrhosis, hepatic failure, immunosuppression, leukaemia/multiple myeloma, lymphoma
ICU admission diagnosis	Categorical: Trauma (reference), Cardiovascular, Gastrointestinal, Respiratory, Genitourinary, Haematologic, Neurologic, Metabolic/Endocrine, Musculoskeletal/skin
Presence of chronic health	Categorical: yes/no
Emergency surgery	Categorical: yes/no
No GCS score due to patient being sedated/paralysed	Categorical: yes/no
Diabetes	Categorical: yes/no
Mechanical ventilation	Categorical: yes/no
Physiological	Heart rate, mean blood pressure, temperature, total respiratory rate, haematocrit, white blood cell count, creatinine, urine output, blood urea nitrogen, sodium, albumin, bilirubin, glucose, PaO_2, acid base abnormalities, GCS score

3.2 External Validation of APACHE IV in HSA ICU

Figure 3.1 illustrates the conceptual model of APACHE IV model for non-CABG patients and clinical variables that were collected within the first day after ICU admission.

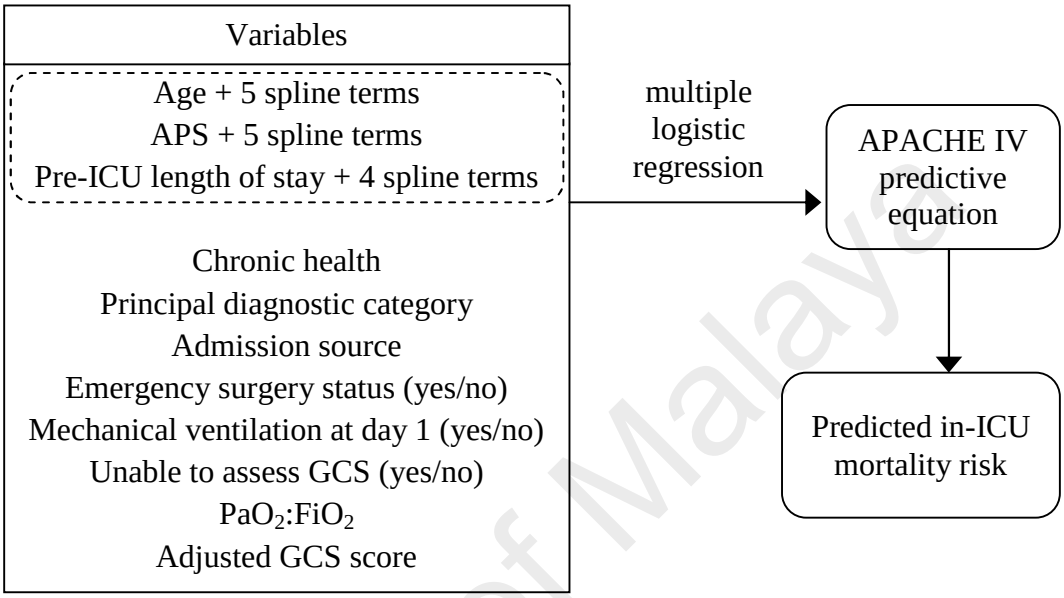


Figure 3.1: APACHE IV Conceptual Model (non-CABG admissions).

Validation of APACHE IV was based on eligible admissions to HSA ICU between the period of 1 January 2009 and 31 December 2009. A comparison between patient characteristics in APACHE IV and HSA ICU was performed in order to have a better understanding of the differences and similarities between the two data sets. Analysis covered demographic aspects such as differences in age, gender make-up, ethnic group and other clinical characteristics.

The Acute Physiology Score (APS) was calculated for each patient who was admitted to HSA ICU throughout the period of study. The APS was computed by combining the scores for the worst physiological variables over the first day of stay in the intensive care unit. Computation of APS for each patient was manually performed using Microsoft® Excel (2007). Patients with incomplete first day APS information

were excluded from analysis so as not to affect the model's predictive accuracy. An imputation method was applied for patients with missing laboratory data, where missing observations were assumed normal and were replaced with midpoint values that were defined in APACHE IV. The midpoint values were used as they were the recommended values in APACHE IV.

The process of fitting the Malaysian ICU data into the multiple logistic regression equation of APACHE IV involved computation of the *logit* term, which was a linear combination of all the variables in APACHE IV. The restricted cubic regression spline terms for age, APS and square root of pre-ICU length of stay were all included in the *logit* term. Details of the regression spline functions and their coefficients in APACHE IV non-CABG model are shown in Appendix E. Probability of death in ICU was calculated using equation (2.3) in Chapter 2, which involved a transformation of the *logit* term.

Model accuracy was evaluated through several measures, i.e. the model’s discrimination, calibration, Standardised Mortality Ratio (SMR) and Brier score. Analysis of model discrimination was performed using MedCalc 10.4 (Medcalc Software, Mariakerke, Belgium), in which estimation of area under receiver operating characteristic curve (AUC) was based on a non-parametric approach by Hanley and McNeil (1982). The model's calibration was assessed through a calibration curve and the Hosmer-Lemeshow's goodness-of-fit test. A *p*-value of less than 5% significance level was used to imply a model's overall lack of fit. The model's SMR was computed by taking the ratio of the mean observed deaths over the mean predicted deaths throughout the duration of study. Model accuracy was also evaluated through the Brier score.

A first level customisation strategy was also applied to improve calibration of APACHE IV in the Malaysian cohort of patients. The approach involved fitting a

simple logistic regression model with observed in-ICU mortality rate being the dependent variable, and the original *logit* term in APACHE IV being the independent variable. The new estimated probability of death for each patient was then calculated from the customised model and calibration for the customised model was evaluated. SPSS 17.0 for Windows (SPSS Inc., Chicago, IL, USA) was used to generate descriptive statistics and perform statistical analysis of variables used in APACHE IV.

3.3 Model Development using Bayesian Markov Chain Monte Carlo approach

The Bayesian Markov Chain Monte Carlo (MCMC) approach was used to develop suitable models for HSA ICU. This approach was applied to identify significant risk predictors and determine the regression coefficients of the proposed models. This study adopted a temporal-split sample approach in model development and validation. Temporal validation can be done by performing a non-random split on a single data set by time, resulting in two periods used for model development and validation (Altman, Vergouwe, Royston, & Moons, 2009). This approach uses data from a pre-determined earlier period for model development, and validation of the model is based on data from a later period in the same cohort of patients. The advantage of this approach is that it allows for non-random variation between the developmental and validation data sets (Moons et al., 2015) and allows a prospective evaluation of the model that is independent of the original data set and developmental process (Altman et al., 2009).

In this study, model development involved all eligible admission data from 1 January 2009 to 31 December 2009. The decision to use all admissions in year 2009 for model development was to allow assessment of seasonality variations or patterns in mortality throughout the whole one year period. This is because seasonal variations in mortality have been well-documented, particularly in respiratory-related admissions (Pendergraft, Stanford, Beasley, Stempel, & McLaughlin, 2005). Validation of the

proposed model was based on data obtained from subsequent admissions between 1 January 2010 and 30 June 2010.

3.3.1 Model Building Strategies - Variable and Weight Selection

One of the most popular model building strategies that is commonly used in ICU prognostic models is the one that was proposed by Hosmer and Lemeshow (2000). This strategy was employed in the development of the Bayesian models. All variables that were in APACHE IV were considered for inclusion in the Bayesian models. Using the variables in APACHE IV as a reference, the aim was to find a parsimonious model that best explains the data and provides accurate predictions. Selection of variables for the Bayesian models was based on the following steps recommended in Hosmer and Lemeshow (2000):

i) Perform a univariate analysis of each possible candidate variable to select the main predictor variables. Variables that were tested to be significant will be fitted into a multivariable logistic regression model.

ii) Fit a new multivariable model that excludes variables that were not significant and compare to the original full model in step (i) using likelihood ratio method.

iii) Include variables that were initially not selected in step (i) in the multivariable model in step (iii). This was to identify variables that may be significant in the presence of other variables, but not important when tested individually. The model at the end of this step was considered as the preliminary main effects model.

iv) Check assumption of linearity in the logit term for continuous variables in the preliminary main effects model.

v) Examine presence of interactions among variables and evaluate whether to include interactions in the multivariable model. This is done by examining all possible combinations of two-way interactions between variables in the model.

Likelihood ratio test is an approach that can be used for model selection and to test the significance of variables in a model (Bagley et al., 2001). To evaluate the significance of a variable, the likelihood ratio test can be applied to compare the -2 log-likelihoods of two nested models, with one model (model without variable) being a subset of the other model (model with variable) (Lewis, Butler, & Gilbert, 2011). The likelihood ratio test is computed as

$$G = -2\,ln\left[\frac{\text{likelihood of model without the variable}}{\text{likelihood of model with the variable}}\right], \tag{3.1}$$

which follows a chi-squared distribution with p degrees of freedom, where p is the difference in the number of parameters between the two models. A p-value < 0.05 implies that there is advantage in including the variable under consideration in the model (Hosmer & Lemeshow, 2000).

In this study, the regression coefficients were estimated using the Bayesian MCMC approach in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Examination of the posterior distributions and their credible intervals can also be used to evaluate significant variables in a model, where a posterior distribution that is far from zero value implies the important contribution of a predictor in a model (Ntzoufras, 2009). In this case, a variable is considered as significant if the estimated credible interval does not contain the value of zero.

In this study, the significance of the estimated regression coefficients for each variable was tested through a combined assessment of frequentist and Bayesian measures, by examining the:

i) likelihood ratio test (comparing likelihoods of model with variable vs. model without variable), and

ii) credible intervals for the posterior means of each variable.

For each univariable model, a variable was considered as candidate for the multivariable model if the *p*-value for the likelihood ratio test was less than 0.25, and if the 75% credible intervals did not contain the value zero. The threshold of 0.25 was chosen based on the argument that a lower *p*-value is often ineffective in screening important variables at the univariate level (Hosmer & Lemeshow, 2000; Mickey & Greenland, 1989).

Variables that satisfied the screening criteria at the univariate level were fitted into various combinations of multivariable models. These variables were then collectively tested for their significance. Linearity assumption for the continuous variables in the models was assessed through Locally Weighted Scatterplot Smoothing (LOWESS) plots (Cleveland, 1979) and non-linear transformation tests (Kay & Little, 1987). Possible interactions between variables were also investigated. In this study, we only considered all possible combinations of two-way interactions between variables. Three-way interactions were not considered in order to reduce model complexity and to prevent overfitting of model.

3.3.2 Model Development using WinBUGS software

Development of the Bayesian models was performed using WinBUGS (Lunn et al., 2000), which is a software that applies the Gibbs Sampling approach (Gelfand & Smith, 1990; Geman & Geman, 1984) in model estimation. Estimation of model parameters is done through an iterative algorithm that takes into consideration information about the prior distribution of the parameters based on past knowledge.

Model development in WinBUGS involved two main parts, i.e. model specification and model inference. Model specification required specification of a likelihood for the response variable, a *logit* expression in the form of a linear combination of potential risk factors, prior distribution for regression parameters, initial values for regression coefficients and input of data that were obtained in the study.

Logistic regression was used to develop the models, where the outcome variable followed a Bernoulli distribution and the *logit* expression was a linear combination of the risk predictors. An additional term was also included in the model in order to account for extra binomial variation. Non-informative priors were used for development of models in this study due to lack of information on the regression parameters. In particular, a weakly informative Gaussian prior distribution with zero mean and a fixed large variance ($\pm^2 = 1000$) was assigned for the regression parameters in the univariable and multivariable models.

The inference part involved steps such as updating the model by running the chain for a fixed number of iterations, making inference on regression parameters, monitoring convergence of chain and obtaining results, summaries and plots. Output posterior summaries that were generated in WinBUGS were used to determine significant variables for the model. In order to monitor convergence of the chains, three multiple parallel chains with different starting points were applied in all simulation work. The univariable models were updated by running the multiple chains for 500,000 iterations each, where the initial 100,000 burn-in samples were discarded from analysis. Simulation runs for the multivariable models were increased to one million iterations, with initial 100,000 burn-in samples.

Model convergence was monitored in WinBUGS through estimated Monte Carlo errors for the posterior means, trace plots and Brooks-Gelman-Rubin (BGR) diagnostic (Brooks & Gelman, 1998). Monte Carlo error (MC error) quantifies the

variability of each estimate caused by Markov Chain simulation (Ntzoufras, 2009). The MC error is used to monitor model convergence, where small values of MC errors indicate better accuracy in parameter estimates. MC errors should also be very much smaller than the standard deviations in a model in order to achieve model convergence. It is recommended that MC error should be less than 5% of the posterior standard deviation (Toft, Innocent, Gettinby, & Reid, 2007).

Other graphical outputs, such as density plots and autocorrelation plots for the posterior distributions of regression coefficients, were used to check for irregularities and chain convergence. Convergence of the MCMC algorithm was also monitored through CODA (Convergence Diagnosis and Output Analysis) package, available in R (R Development Core Team, 2013). Analysis of CODA in R involved four diagnostic tests, i.e. Geweke Convergence Diagnostic (Geweke, 1992), Gelman and Rubin Diagnostic (Gelman & Rubin, 1992), Raftery and Lewis Convergence Diagnostic (Raftery & Lewis, 1992) and Heidelberger and Welch Stationarity and Interval Halfwidth Tests (Heidelberger & Welch, 1983).

The model's goodness-of-fit was evaluated through a measure known as Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), where a model with a lower DIC value was considered to have better fit. As a guide, Spiegelhalter et al. (2002) recommended that models with differences of DIC greater than 3 implies that the model with smaller DIC is preferred, and a DIC difference of greater than 10 rules out the model with the higher DIC.

3.3.3 Proposed types of Bayesian models

Several types of approaches were explored in the construction of the multivariable models, especially in the modelling of the physiological variables. The different types of models that were considered are summarised as follows:-

i) Type W: Main effects model with Acute Physiology Score (APS) variable (scores assigned to worst values of physiological observations within the first day of ICU admission),

ii) Type M: Main effects model with APS.mean variable (scores assigned to mean values of physiological observations within the first day of ICU admission),

iii) Type P: Main effects model with dichotomous abnormal physiological variables (yes/no),

iv) Type A: Main effects model with frequency of abnormal physiological variables,

v) Type F: Main effects model using factor analysis approach for worst physiological variables.

In the first approach, the Acute Physiology Score (APS) and other variables that satisfied the screening criteria at the univariate level were fitted into several combinations of multivariable models. These variables in the multivariable models were then collectively tested for their significance in order to identify the main effects models. Possible interactions between variables were tested for their significance in the multivariable models and the linearity assumption between continuous variables and the outcome variable was also checked. The *logit* term for the first type of model (Type W) was characterised as

$$logit_{Type\,W} = \beta_0 + \sum_{k=1}^{k=n-1} \beta_k x_k + \beta_{\mathrm{APS}} \cdot \mathrm{APS} + \sum_{j \neq k} \sum \beta_{jk} x_j x_k \,, \qquad (3.2)$$

where β_0 was the constant term, β_k denoted the regression coefficient for variable x_k, β_{jk} was the regression coefficient for possible interaction term between variables j and k $(j \neq k)$ and β_{APS} represented the regression coefficient for APS variable in a multivariable model with n variables.

The effect of using mean values of physiological observations instead of the worst values within the first day of ICU admission was explored in the second type of models (Type M). All variables that were used in the first type of models were also applied in the development of the Type M models. The only difference was that the APS variable was substituted with a new variable known as APS.mean. Calculation of APS.mean was based on scores that were assigned to the mean values of the physiological variables within the first day of ICU admission.

One of the important innovations in this study was to explore alternative approaches to substitute the use of APS variable in the predictive models. As replacement for the APS variable, the worst values for each physiological variable within the first day of ICU admission were dichotomously coded as normal/abnormal and directly included into the multivariable model. These physiological variables were included together with other important variables from the main effects model in the third type of models. Classification of abnormality was based on the APACHE IV scoring rule, where values within the normal range were specifically those with zero scores. The *logit* term for the third type of model (Type P) is given as

$$logit_{Type\,P} = \beta_0 + \sum_{k=1}^{k=n-p} \beta_k x_k + \sum_{k=n-p+1}^{n} \beta_k x_k, \tag{3.3}$$

where β_0 is the constant term, β_k is the regression coefficient for variable x_k, $k = 1,..., n$. The $logit_{Type\ P}$ term consists of two parts. The first part is the linear combination of $(n - p)$ variables from the main effects model, whereas the second part is the linear combination of p dichotomous physiological variables.

The construction of the fourth type of model (Type A) required assessment of the frequency of abnormal physiological observations throughout the first day of ICU admission. Classification of abnormality for the physiological variables was based on the ranges defined in APACHE IV. Two types of approaches were explored in the

calculation of frequency of abnormal physiological observations. In the first approach, the worst value for each physiological variable within the first day of ICU admission was classified as normal/abnormal. A new variable known as *physio.ooc*, which represented the total percentage of abnormal physiological observations for a specific patient, was derived as

$$physio.ooc = \frac{\sum \text{number of abnormal physiological variables}}{\sum \text{number of physiological variables}}. \tag{3.4}$$

The *logit* term for the corresponding model is

$$logit_{Type\ A} = \beta_0 + \sum_{k=1}^{n-1} \beta_k x_k + \beta_{physio.ooc} \cdot physio.ooc, \tag{3.5}$$

where β_0 is the constant term, $\beta_{physio.ooc}$ is the regression coefficient for the *physio.ooc* variable and β_k denotes the regression coefficient for other variables that were in the main effects model x_k, $k = 1,..., n-1$.

The second approach used in the construction of Type A models involved direct incorporation of the frequency of abnormal observations for each individual physiological variable into the multivariable model. For each of the physiological variable, the frequency of abnormalities observed within the first day of ICU admission is given as

$$ooc.var(j) = \frac{\sum \text{number of abnormal observations for variable } j}{\sum \text{total observations for variable } j}, \tag{3.6}$$

where $j = 1,..., p$ for p number of physiological variables. The *logit* term for the multivariable model is

$$logit_{Type\ A} = \beta_0 + \sum_{k=1}^{n-p} \beta_k x_k + \sum_{j=1}^{p} \beta_{ooc.var(j)} \cdot ooc.var(j), \tag{3.7}$$

where $_0$ is the constant term, $_{ooc.var(j)}$ is the regression coefficient for *ooc.var(j)* and $_k$ represents the regression coefficient for other variables that were in the main effects model.

Factor analysis can be used to identify latent (unobserved) variables or factors that describe the relationship or patterns in a group of observed variables (Yong & Pearce, 2013). The principle behind factor analysis is that observed variables are assumed to be affected by common underlying factors (latent variables) and unique factors. The common factors (latent variables) are unobserved variables that are indirectly measured by observed variables. The advantage of of factor analysis is that it is able to assess intercorrelation of variables in a data set and allows better understanding of how variables are inter-related (Warner, 2012). This is beneficial especially in determining relationships between physiological variables in a prognostic model. Based on this motivation, in the development of the fifth type of model (Type F), factor analysis was explored as a data reduction approach to determine underlying common factors that describes the physiological components of the model. Factor analysis allowed the p physiological variables $x_{1i}, x_{2i}, ..., x_{pi}$ for patient i $(i = 1, ..., n)$ to be expressed as linear functions of common factors $F_{1i}, F_{2i}, ..., F_{mi}$ $(m < p)$ and unique error terms $U_{1i}, U_{2i}, ..., U_{pi}$, i.e.

$$x_{ji} = a_{j1}F_{1i} + a_{j2}F_{2i} + ... + a_{jm}F_{mi} + a_{ji}U_{ji}, \qquad i = 1, ..., n, \quad j = 1, ..., p. \tag{3.8}$$

An advantage of the factor analysis approach was that it reduced the large number of physiological variables to only several common unobservable factors, without loss of information (Yong & Pearce, 2013). These common factors, also known as latent factors, described the correlation between physiological variables and classified them into groups that have factors in common. Due to differences in the measurement units of the physiological variables, standardisation of the variables was

performed in order to achieve meaningful interpretation. This resulted in the following expression:

$$z_{ji} = a_{j1}F_{1i} + a_{j2}F_{2i} + ... + a_{jm}F_{mi}, \qquad i = 1,...,n, \quad j = 1,..., p, \tag{3.9}$$

where z_{ji} is the standardised variable. The number of factors to be retained in a model can be determined through either Eigen values or scree test (Costello & Osbourne, 2005). In this study, Eigen values and scree plots were used to determine the number of factors to be retained, where factors with Eigen values greater than 1.0 were retained. Factor analysis approach was performed using SPSS 17.0 for Windows (SPSS Inc. Chicago, IL, USA).

3.4 Model Assessment

The overall predictive performance of these Bayesian models was assessed through the Standardised Mortality Ratio (SMR) and Brier score. SMR was computed as the ratio of the mean observed deaths over the mean predicted deaths. A ratio of 1.0 indicated that the expected and observed hospital death rates were the same. A ratio that was greater than 1.0 indicated that the observed mortality was higher than the expected mortality, while a value that was lower than 1.0 suggested otherwise. A lower Brier score was considered to indicate better model fit.

Model discrimination was evaluated through area under receiver operating characteristic curve (AUC) using MedCalc 10.4 (MedCalc Software, Mariakerke, Belgium). Estimation of AUC was based on the Hanley and McNeil (1982) non-parametric approach. In theory, perfect discrimination in a model is achieved when AUC = 1.0, where predicted outcomes will be the same as the observed outcomes for all patients. Lower values of AUC indicated less discriminatory power, whereas higher AUC values reflected better discrimination. Model discrimination was considered good if AUC > 0.8 in this study.

Model calibration was assessed through calibration curves and the Hosmer-Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1980). The Hosmer-Lemeshow test required the construction of ten groups for different categories of patients, where their predicted and observed mortality probabilities were compared within each subgroup. Patients were sorted according to their predicted mortality probabilities and subsequently divided into ten subgroups with approximately equal size. The Hosmer-Lemeshow test statistic was then computed using equation (2.13) in Section 2.9.2. The overall model calibration was considered good if the Hosmer-Lemeshow goodness-of-fit test was non-significant at p-value > 0.05. Calibration curves were also plotted for each model to compare the mean predicted probabilities against the mean observed outcomes for each subgroup of patients. The model's goodness-of-fit was also evaluated through Deviance Information Criterion (DIC) values that were estimated from samples that were generated in the Bayesian MCMC simulation. Comparison of the performance of competing models was done through assessment of DIC values. A model with a smaller DIC value was considered to have better fit. The performances of the different types of Bayesian models that were proposed in this study were compared using the above-mentioned criteria. The model with the best overall performance was then chosen as the final proposed model in this study. For comparison purpose, the estimates and standard errors of regression coefficients in the best model were also obtained through the frequentist (maximum likelihood estimation) method using S-PLUS ver 8.0 (Insightful Corp., Seattle, U.S.A). The performance of the frequentist model was then compared against the Bayesian model.

3.5 Mortality Prediction using Decision Tree approach

The decision tree approach was explored as an alternative method to predict in-ICU mortality risk in this study. A decision tree was constructed for only the "best" model

among the Bayesian models proposed in Section 3.3.3. All of the variables in the "best" model were input into the decision tree model, with risk of death in ICU being the outcome variable. Construction of the decision tree was based on a set of training data, based on 916 HSA ICU admissions between 1 January 2009 and 31 December 2009. Validation was performed using a different data set, comprising of 195 admissions between 1 January 2010 and 30 June 2010. The training and validation data sets were the same ones used in development of the Bayesian models.

Analysis and construction of the decision tree was performed using SPSS 17.0 for Windows (SPSS Inc. Chicago, IL, USA). The CHAID algorithm, which was available in the software package, was chosen for this analysis. This approach was considered more appropriate and relevant because of its ability to always select independent predictors that have the strongest association with the outcome variable at each step (Badriyah et al., 2012). Moreover, the algorithm does not require pruning to be done and does not generate overly large trees, as compared to the CART algorithm (Rokach & Maimon, 2008). The comparatively large data set and huge number of predictors in this study was also considered conducive for application of the CHAID algorithm.

The advantage of the decision tree approach is that it allows easy interpretation and visual trace of paths along the tree (Song & Lu, 2015). Results that are generated in the terminal nodes of the decision tree provide the predicted risk of mortality for patients presenting various specific conditions. In this analysis, the predicted results that were obtained from the decision tree were compared against the actual values. The performance of the decision tree approach was also assessed using standard measures such as SMR, AUC and the Hosmer-Lemeshow goodness-of-fit test. These measures were then compared against the results obtained using the Bayesian MCMC approach for the same model.

CHAPTER 4: ANALYSIS AND FINDINGS

The demographic and clinical characteristics of patients who were admitted to HSA throughout the period of study are described in Section 4.1. Findings on the external validation of APACHE IV in HSA ICU are presented in Section 4.2. Results that were obtained for the proposed models that were developed using Bayesian MCMC approach are discussed in Section 4.3. Analysis of the results obtained using logistic regression decision tree approach is given in Section 4.4.

4.1 Patient characteristics

Of the 1,084 patients who were admitted to HSA ICU between 1 January 2009 and 31 December 2009, 168 admissions failed to meet the exclusion criteria defined in Chapter 4.2.2. A total of 916 eligible admissions were used for validation of APACHE IV model in HSA ICU, as well as, for the developmental stage of Bayesian MCMC models in this study. There were 200 admissions in HSA ICU between 1 January 2010 and 30 June 2010. Application of the exclusion rules resulted in 5 patients being removed, leaving only 195 admissions for the validation stage of Bayesian MCMC models.

A comparison of the differences in demographic and clinical characteristics for HSA ICU admissions between year 2009 and the first half of 2010 is shown in Table 4.1. Patients who were admitted to HSA ICU were predominantly male (approximately 60%) and the majority of patients were on mechanical ventilation. The Malay patients formed slightly more than half of the total admissions, followed by Chinese, Indian and patients from other ethnic groups. Overall, approximately 48% of patients were admitted directly from the hospital's ward or recovery room, whereas 40% of them were transferred to the ICU from the operating room or emergency room. The remaining 12% were made up of patients who were transferred to the ICU from another special care

unit in the same hospital. Transfer cases from other hospitals were removed from analysis.

Table 4.1: Characteristics of HSA ICU admissions.

Patient characteristics	Stage 1[#] (Developmental data set)	Stage 2[*] (Validation data set)	Overall
Total patients	916	195	1,111
Age (mean ± SD, in years)	43.4 ±17.6	43.6 ± 18.5	43.5 ± 17.7
Acute Physiology Score, APS (mean ± SD)	69.6 ± 31.9	63.3 ± 33.1	68.5 ± 32.2
Male (%)	60.6	61.5	60.8
Ethnicity (%)			
Malay	56.4	53.3	55.9
Chinese	24.1	27.2	24.7
Indian	10.7	11.3	10.8
Others	8.7	8.2	8.6
ICU admission source (%)			
Floor	47.3	49.7	47.7
Other special care unit	12.2	10.8	12.0
Operating room	40.5	39.5	40.3
Emergency surgery (%)	36.6	35.4	36.4
Pre ICU length of stay (mean ± SD, in days)	1.1 ± 2.3	0.8 ± 1.7	1.1 ± 2.2
Mechanically ventilated (%)	83.0	86.7	83.6
Unable to obtain Glasgow Coma Scale (GCS) score (%)	23.1	35.9	25.4
Dead in ICU (%)	18.8	16.9	18.5
With at least one comorbidities (%)	3.7	4.1	3.8
Diabetes (%)	20.1	21.5	20.3
Disease categories (%)			
Trauma	20.6	19.0	20.3
Cardiovascular	22.3	19.0	21.7
Respiratory	18.2	20.5	18.6
Neurologic	17.1	16.4	17.0
Gastrointestinal	11.1	9.2	10.8
Genitourinary	7.1	9.2	7.5
Metabolic/endocrine	2.6	1.5	2.4
Musculoskeletal/skin	0.5	1.5	0.7
Haematologic	0.3	3.6	0.9

Data collected from 1 January 2009 to 31 December 2009

* Data collected from 1 January 2010 to 30 June 2010

Admissions to HSA ICU were almost equally divided between non-operative and post-operative patients. Figure 4.1 and Figure 4.2 illustrate the comparison of principal admission diagnostic categories for non-operative and post-operative patients

in year 2009 and the first half of 2010 respectively. Most of the post-operative patients were admitted due to trauma, whereas cardiovascular and respiratory diseases were the main cause of ICU admission for the majority of non-operative patients. There were only few patients who were admitted for musculoskeletal/skin and haematologic diseases. Post coronary artery bypass graft patients were treated in a separate unit and were not included in the study.

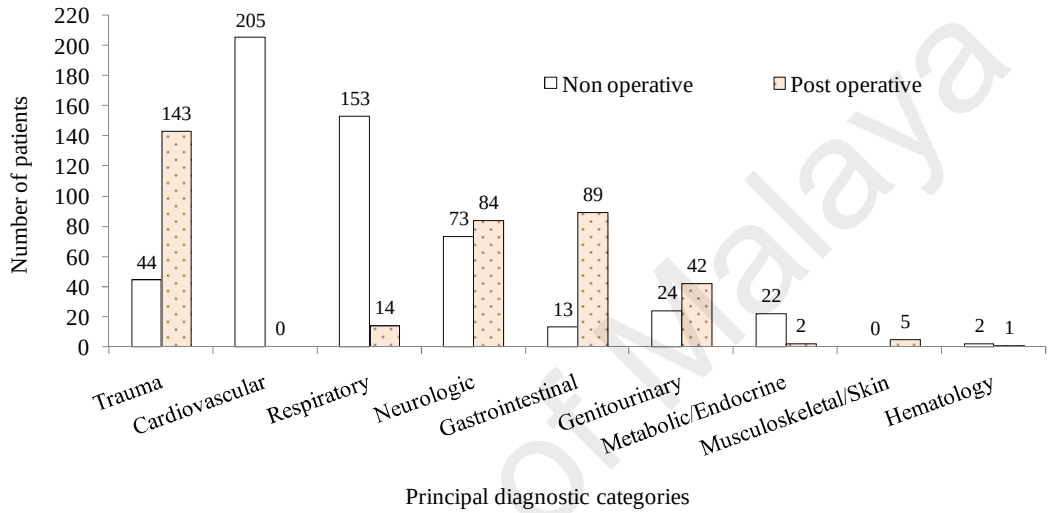


Figure 4.1: Disease categories for admissions to HSA ICU in 2009.

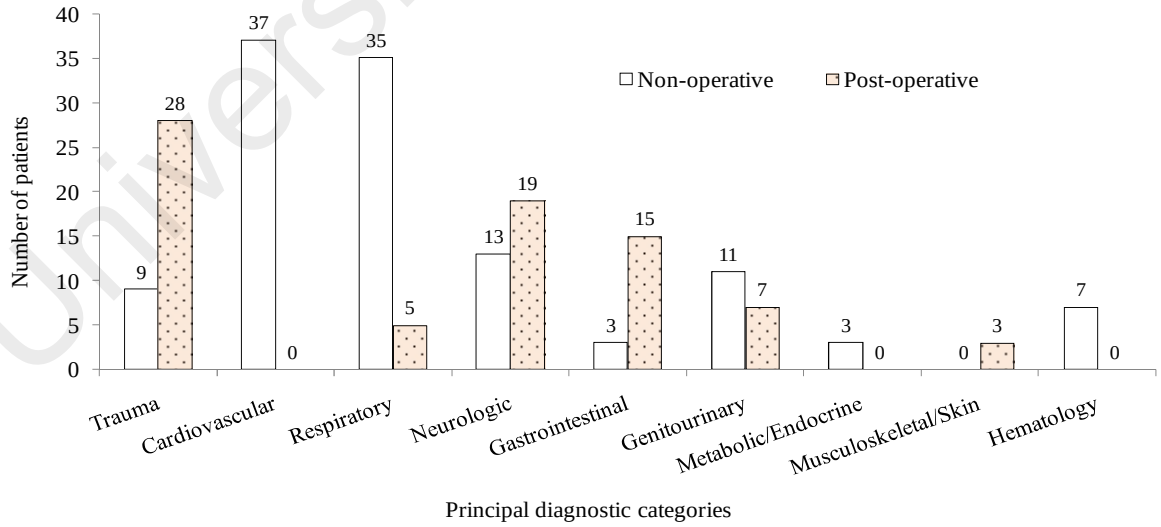


Figure 4.2: Disease categories for admissions to HSA ICU in first half of 2010.

The cohort of patients in HSA ICU for the whole period of study generally had a low comorbidity load, with only 42 patients (3.8%) who reported that they had at least

one existing comorbidities defined in APACHE IV (see Table 4.1). About 45% of these patients were suffering from immunodeficiency disorders (immunosuppression), followed by metastatic cancer (19%), cirrhosis (14%), AIDS (7%) and leukaemia/multiple myeloma (7%). There were only two patients with lymphoma and only one patient with hepatic failure (see Figure 4.3).

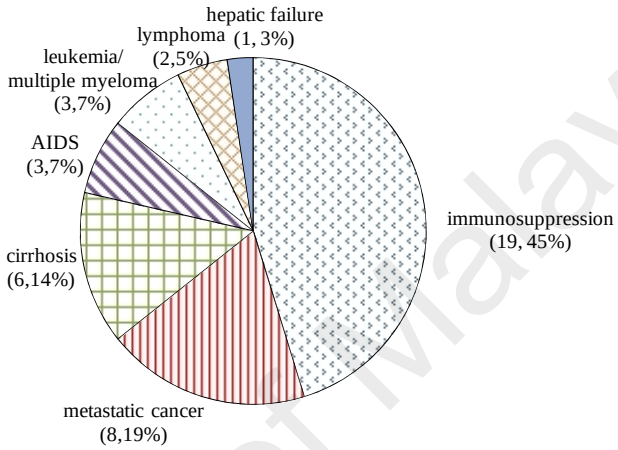


Figure 4.3: Percentage and number of HSA ICU admissions with different types of comorbidities between 1 January 2009 and 30 June 2010.

Throughout the whole period of study from 1 January 2009 to 30 June 2010, a total of 226 patients (20.3%) revealed that they had diabetes mellitus. This high figure was consistent with the findings obtained by Letchumanan et al. (2010), which reported that the overall prevalence of diabetes in Malaysia (11.6%) was higher compared to other regions in the world. Further analysis revealed that the prevalence of diabetes in HSA ICU was higher in older patients, aged 50 years and above (see Figure 4.4). There was also not much of a difference in the prevalence of diabetes among the three major ethnic groups (Malay, Chinese and Indian), as shown in Figure 4.5.

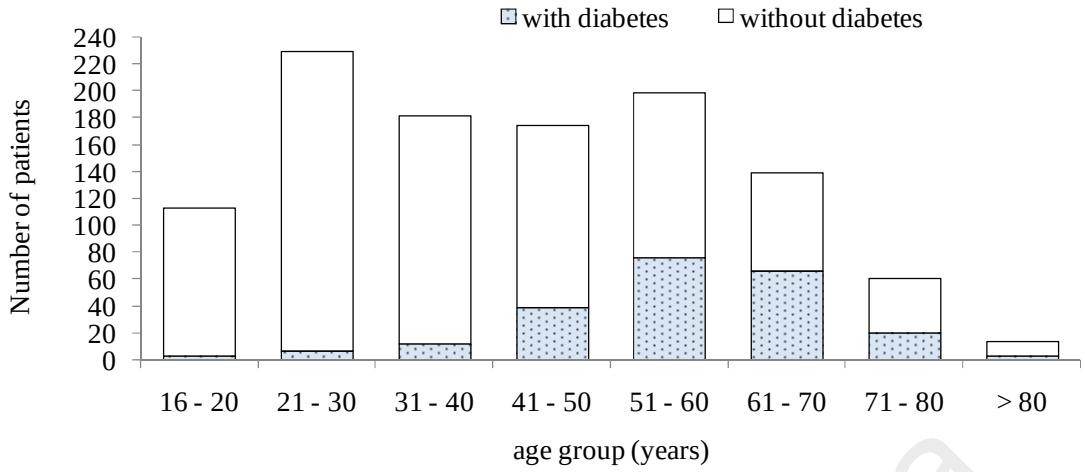


Figure 4.4: Number of HSA ICU patients with and without diabetes for different age groups between 1 January 2009 and 30 June 2010.

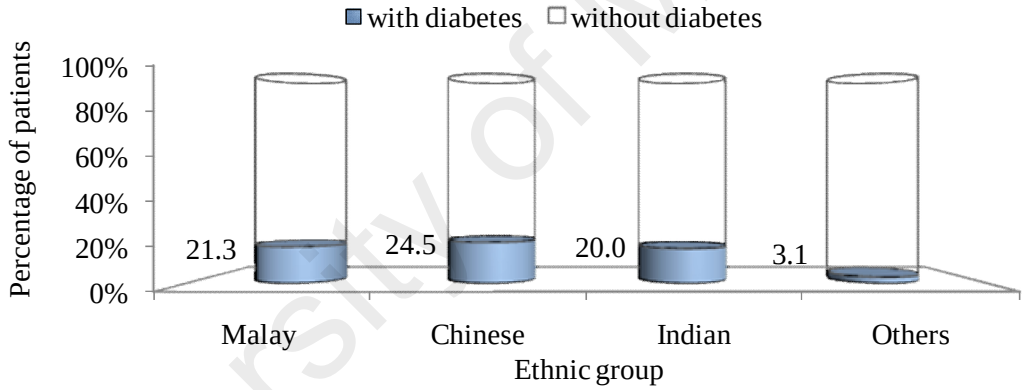


Figure 4.5: Percentage of HSA ICU patients with and without diabetes for different ethnic groups between 1 January 2009 and 30 June 2010.

Admissions to HSA ICU throughout the period of study generally consisted of a younger set of patients, with an overall mean age of 43.5 years (±17.7 years). The age distribution for HSA ICU was positively skewed and the mean was significantly affected by a high proportion of younger patients (see Figure 4.6).

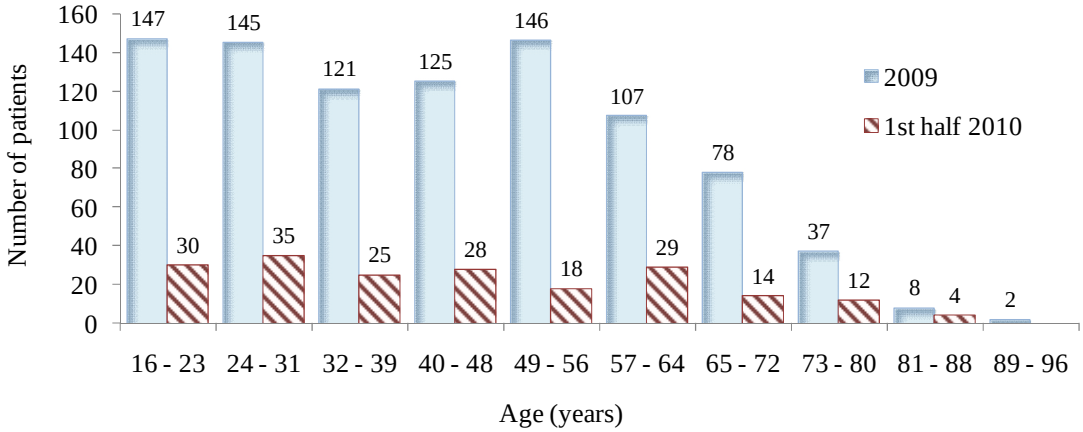


Figure 4.6: Histogram of age distribution for admissions to HSA ICU.

The majority of younger patients (below age 30 years old) were admitted due to trauma-related illnesses, whereas patients between 30 to 50 years old were mostly admitted because of cardiovascular and neurologic diseases. A large percentage of middle age patients (in their 50s and 60s) were admitted due to cardiovascular and respiratory ailments, while older admissions (beyond 70 years old) were mostly due to gastrointestinal problems (see Figure 4.7).

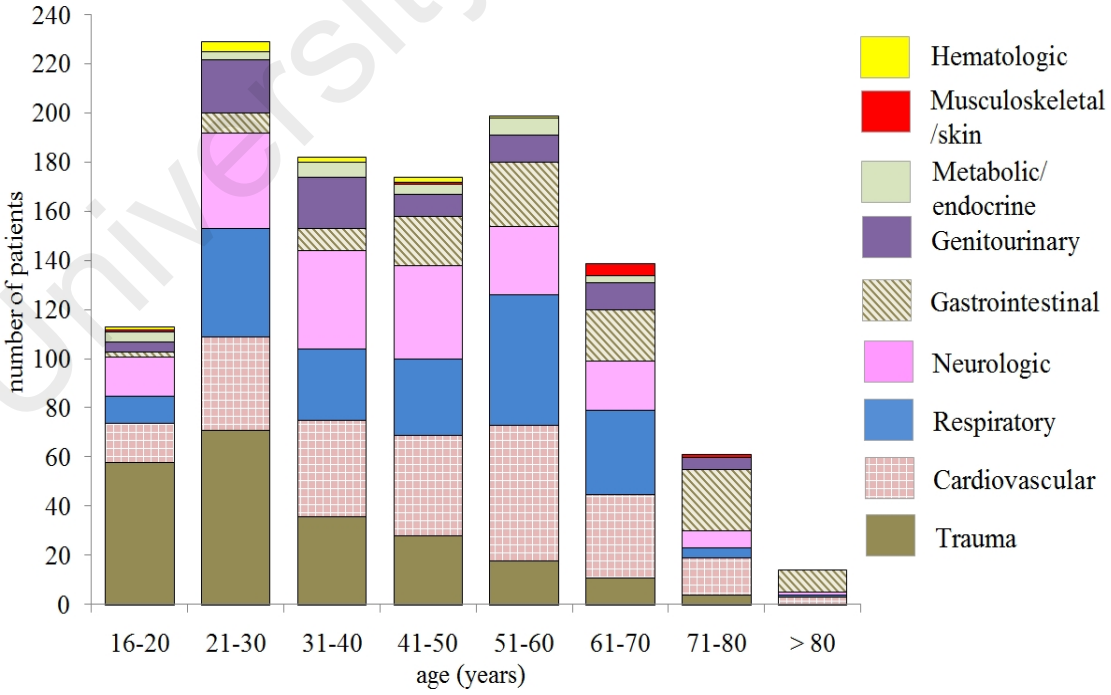


Figure 4.7: Principal admission diagnosis according to age groups for admissions to HSA ICU between 1 January 2009 and 30 June 2010.

The number of in-ICU deaths at HSA ICU between 1 January 2009 and 30 June 2010 was 205 (18.5%). Out of the 916 eligible admissions in year 2009, 172 (18.8%) died in ICU. The first half of 2010 registered 33 (16.9%) in-ICU deaths out of 195 eligible admissions. Figure 4.8 illustrates the number of deaths in HSA ICU by ethnic groups. There was not much of a difference among the ethnic groups in the percentage of patients who died in ICU. The Indian patients registered the lowest percentage among the ethnic groups. However, the number of Indian patients who were admitted to HSA ICU was also the lowest among the ethnic groups.

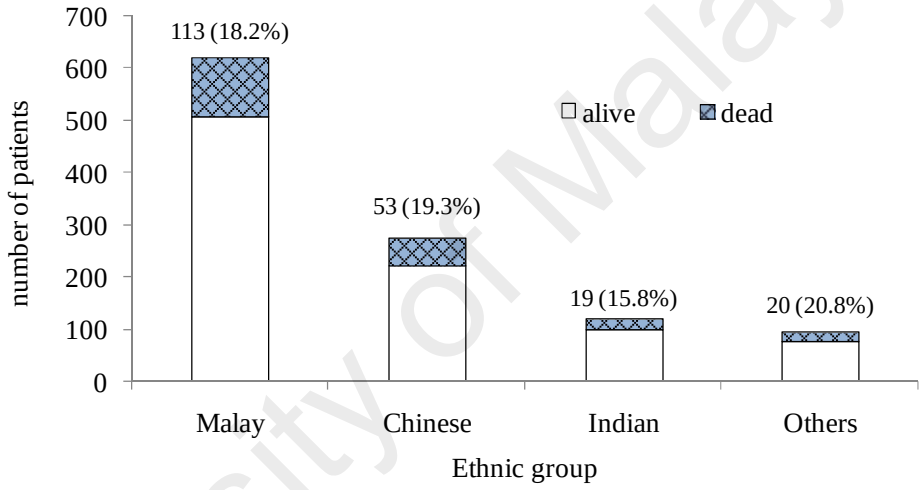


Figure 4.8: Number of deaths in HSA ICU according to ethnic groups between 1 January 2009 and 30 June 2010.

Patients who were admitted to HSA ICU generally exhibited greater degree of severity of illness. The overall mean of first day APS for admissions to HSA ICU in year 2009 (69.6) and the first half of 2010 (63.3) were observed to be rather high values. The first day APS distribution for HSA ICU admissions in year 2009 and the first half of 2010 are shown in Figures 4.9 and 4.10 respectively. The APS range for HSA ICU admissions in year 2009 was between 11 and 171. Admissions in the first half of 2010 registered a minimum APS of 16 and a maximum APS of 164. Test for normality suggested evidence that APS was not normally distributed, where the majority of

patients had APS values between 41 and 50. There were also isolated cases of patients with extremely high APS values exceeding 140.

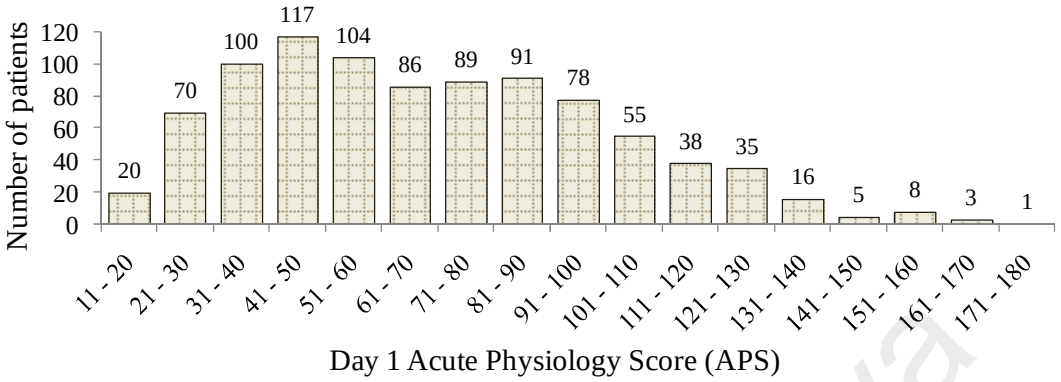


Figure 4.9: Day 1 APS for HSA ICU admissions in 2009.

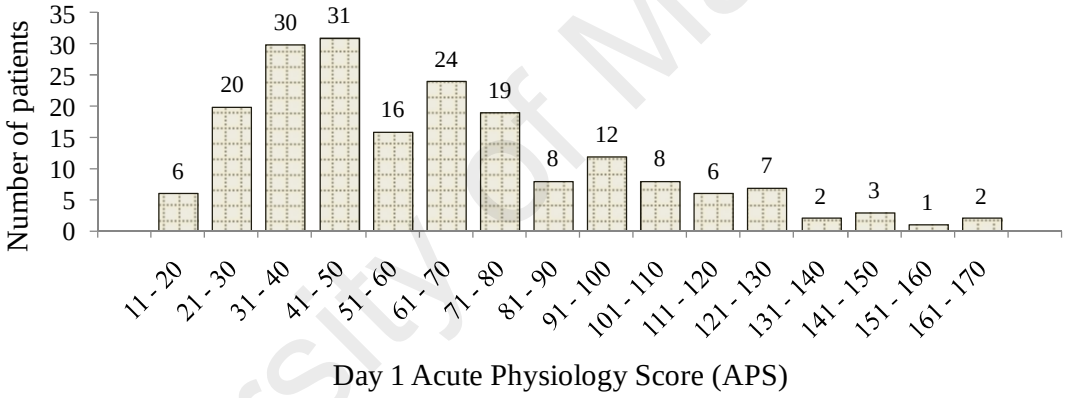


Figure 4.10: Day 1 APS for HSA ICU admissions in first half of 2010.

In theory, patients who have higher APS values are often associated with higher risks of death. An analysis was performed to determine the relationship between APS and outcome of patients upon ICU discharge. Boxplot comparison of APS between groups of patients who died and those who were alive for the year 2009 and the first half of 2010 are shown in Figure 4.11. On the whole, the median APS for the group of patients who died in ICU was significantly higher compared to the median APS for the group that survived. Some outliers were observed in the APS for the group of patients

who were still alive upon ICU discharge. Despite having extremely high APS values, these few patients defied the odds of dying and miraculously survived.

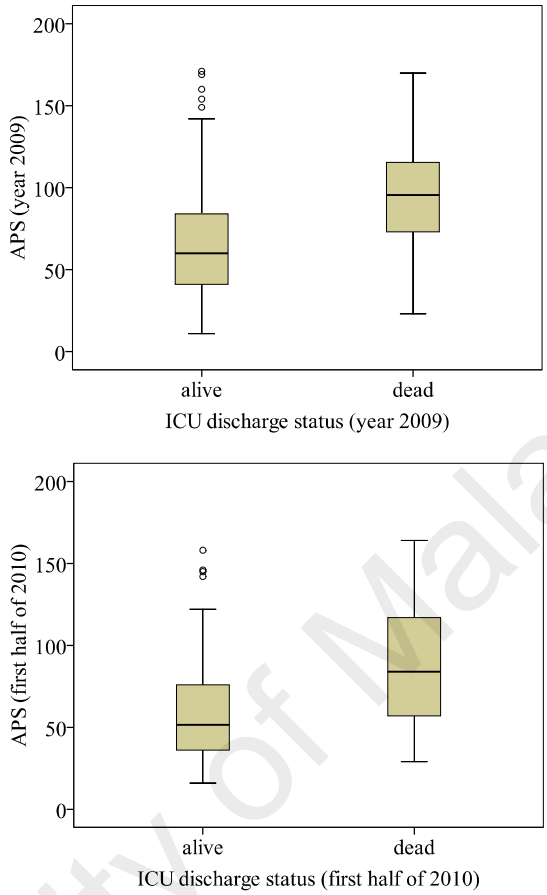


Figure 4.11: Boxplot comparison of APS for HSA ICU patients who were dead and alive upon ICU discharge for year 2009 and the first half of 2010.

A simple linear regression test was performed to analyse the relationship between age and APS variables using data that were obtained between 1 January 2009 and 31 December 2009. The scatter plot in Figure 4.12 did not reflect any positive trend between age and APS of patients who were admitted throughout year 2009. Patients with high APS were not necessarily older since there were also many younger patients with high APS values. Results that were obtained indicated a very weak positive linear relationship between age and APS, with a correlation coefficient, r = 0.135 (see Table 4.2).

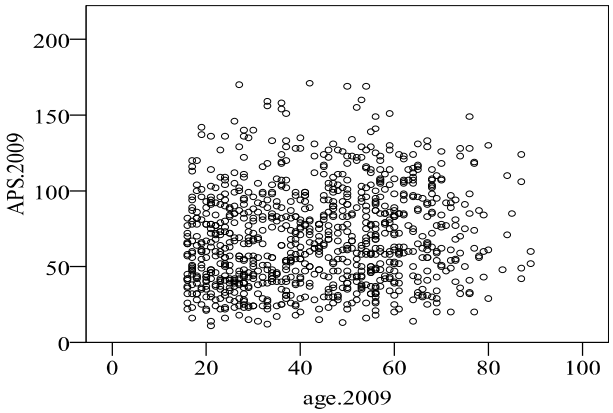


Figure 4.12: Scatter plot of age versus APS for HSA ICU admissions in year 2009.

Table 4.2: Output summary of linear regression test between age and APS.

Model Summary

R	R Square	Adjusted R Square	Std.Error of the Estimate
0.135	0.018	0.017	31.666

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	16892.801	1	16892.801	16.847	0.000
Residual	916501.963	914	1002.737		
Total	933394.737	915			

It is commonly believed that patients with a longer length of stay in the hospital prior to ICU admission usually have higher risks of deaths (Nahra, Schorr, & Gerber, 2005). However, the data that were obtained in this study suggested no significant differences in the means of pre-ICU length of stay between patients who died and those who were alive upon discharge from HSA ICU (see Figure 4.13). This suggested the possibility that pre-ICU length of stay was probably not a significant predictor of in-ICU mortality in the context of this Malaysian study (p-value = 0.71, two sample t-test). The use of pre-ICU length of stay variable is more related to the practice and management of ICU, especially as a determinant in allocation of ICU cost and resources. Figure 4.14 depicts the histogram of pre-ICU length of stay for admissions between 1 January 2009 and 30 June 2010. A square-root transformation was applied to

the pre-ICU length of stay variable as the initial data set was positively skewed. After transformation, the distribution of pre-ICU length of stay appeared to be still positively skewed and non-normal, as indicated by the summary statistics in Table 4.3.

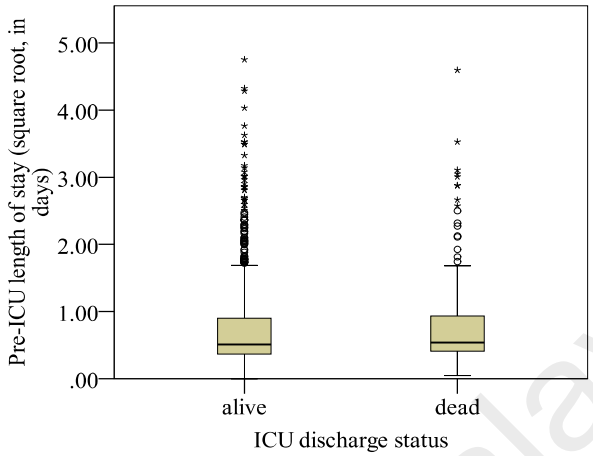


Figure 4.13: Comparison of Pre-ICU length of stay (square root days) between patients who were alive and dead upon ICU discharge from 1 January 2009 to 30 June 2010.

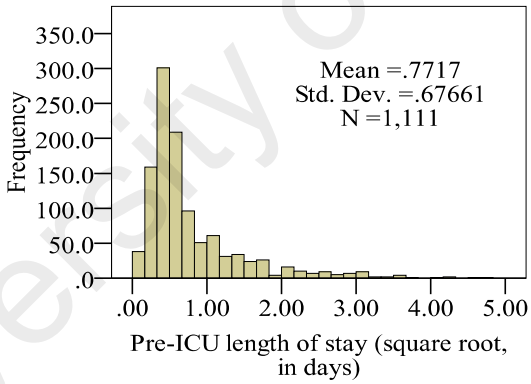


Figure 4.14: Histogram of Pre-ICU length of stay (in square root days).

Table 4.3: Summary statistics of Pre-ICU length of stay variable.

N	Min.	Max.	Mean		Standard Deviation	Skewness		Kurtosis	
Stat.	Stat.	Stat.	Stat.	Std. Error	Stat.	Stat.	Std. Error	Stat.	Std. Error
1111	0.00	4.75	0.7717	0.02030	0.67661	2.317	0.073	6.371	0.147

4.2 Performance of APACHE IV in HSA ICU

4.2.1 Comparison between HSA ICU and APACHE IV data sets

Results in Table 4.4 suggested significant differences in demographic and clinical characteristics between HSA ICU admissions for the year of 2009 and APACHE IV (developmental sample). Admissions to HSA ICU generally recorded a higher percentage of male patients (60.6%), compared to the APACHE IV developmental sample (54.2%). The ethnic compositions in the Malaysian ICU comprised four categories (Malay, Chinese, Indian and Others), whereas there were only two categories of race defined in APACHE IV (white and non-white). The percentage of post-operative admissions in HSA ICU (49.3%) was higher compared to APACHE IV (30.9%). A high proportion of HSA ICU post-operative admissions were emergency surgery cases (36.6%), as compared to a much lower percentage of emergency surgery cases in APACHE IV (5.7%). More than 80% of patients who were admitted to HSA ICU were on mechanical ventilation, as compared to only 35.1% in APACHE IV.

The mean age of patients in HSA ICU at 43.44 years was much lower than the corresponding mean of 61.51 years in APACHE IV. Only a small percentage of HSA ICU patients (3.7%) disclosed that they have at least one of the seven comorbidities defined in APACHE IV. The reasons for this relatively low figure could be due to patients who deliberately withhold important information about their underlying conditions due to fear of stigma, or could not provide accurate information due to lack of awareness of their previous medical conditions. The mean APS for ICU Day 1 admissions to HSA ICU (69.59) was also significantly higher than the mean APS for APACHE IV (38.83). The overall percentage of deaths in HSA ICU for the period of study was also observed to be higher than the corresponding mortality rate in APACHE IV.

Table 4.4: Comparison of patient characteristics between HSA ICU (1 January 2009 to 31 December 2009) and APACHE IV developmental sample.

Characteristics	HSA ICU (n = 916)	APACHE IV (n = 66,270)
Gender (% male)	60.6	54.2
Ethnic group		
White	-	69.3
Malay	56.4	-
Chinese	24.1	-
Indian	10.7	-
Others	8.7	-
Mean age (years)	43.44 ± 0.58	61.51 ± 0.07
Mean APS	69.59 ± 1.06	38.83 ± 0.10
Mean pre ICU length of stay (square root days)	0.794 ± 0.023	0.786 ± 0.004
Died in ICU (%)	18.8	13.6
With comorbidities (%)	3.7	10.3
Emergency surgery (%)	36.6	5.7
Postoperative patient (%)	49.2	30.9
Ventilated on ICU Day 1 (%)	83.0	35.1
Unable to assess GCS (%)	23.1	8.0

4.2.2 Validation of APACHE IV model in HSA ICU

The performance of APACHE IV in HSA ICU was assessed through several indicators. A comparison of performance indicators between HSA ICU and APACHE IV is shown in Table 4.5. The observed in-ICU mortality rate for HSA ICU was much lower than the overall predicted in-ICU mortality rate, with an approximate difference of 9%. Application of APACHE IV in HSA ICU also resulted in an overall SMR of 0.668, which was much lower than the ideal value of 1.0. On the whole, APACHE IV exhibited acceptable discrimination when applied to the HSA ICU cohort of patients, with an area under receiver operating characteristic curve (AUC) value of 0.78 (see Figure 4.15 and Table 4.6). However, the model's calibration in HSA ICU was observed to be poor, as indicated by the Hosmer-Lemeshow goodness-of-fit $\hat{C}$ statistic (Table 4.5) and calibration curve (Figure 4.16).

Table 4.5: Performance comparison between HSA ICU and APACHE IV

Performance indicators	HSA ICU (n = 916)	APACHE IV (n = 66,270)
Observed in-ICU mortality rate (%)	18.78	13.51
Predicted in-ICU mortality rate (%)	28.11	13.55
Standardised Mortality Ratio (SMR)	0.668	0.997
Area under ROC curve (AUC)	0.78	0.88
Hosmer-Lemeshow statistic (p-value)	113 (p<0.0001)	16.8 (p=0.08)

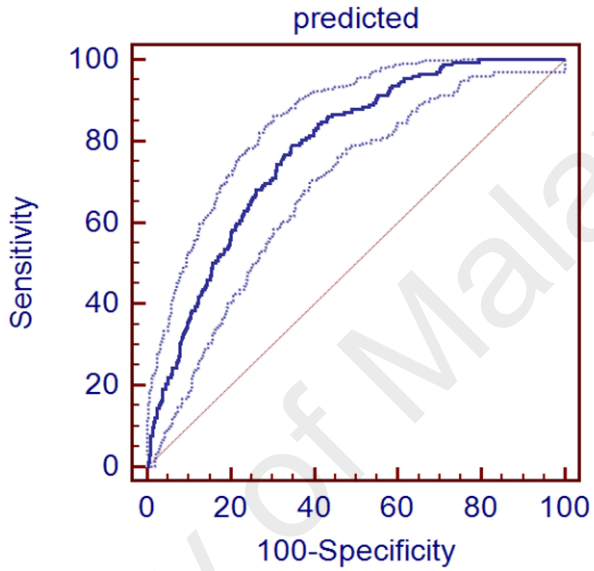


Figure 4.15: Receiver operating characteristic curve for validation of APACHE IV in HSA ICU.

Table 4.6: Area under receiver operating characteristic curve summary results for validation of APACHE IV in HSA ICU.

Variable	predicted
Classification variable	actual
Sample size	916
Positive group: actual = 1	172
Negative group: actual = 0	744
Disease prevalence (%)	Unknown
Area under the ROC curve (AUC)	0.780
Standard Error[a]	0.0219
95% Confidence Interval[b]	0.751 to 0.806
z statistic	12.772
Significance level P (Area=0.5)	<0.0001

[a] Hanley & McNeil, 1982
[b] Binomial exact

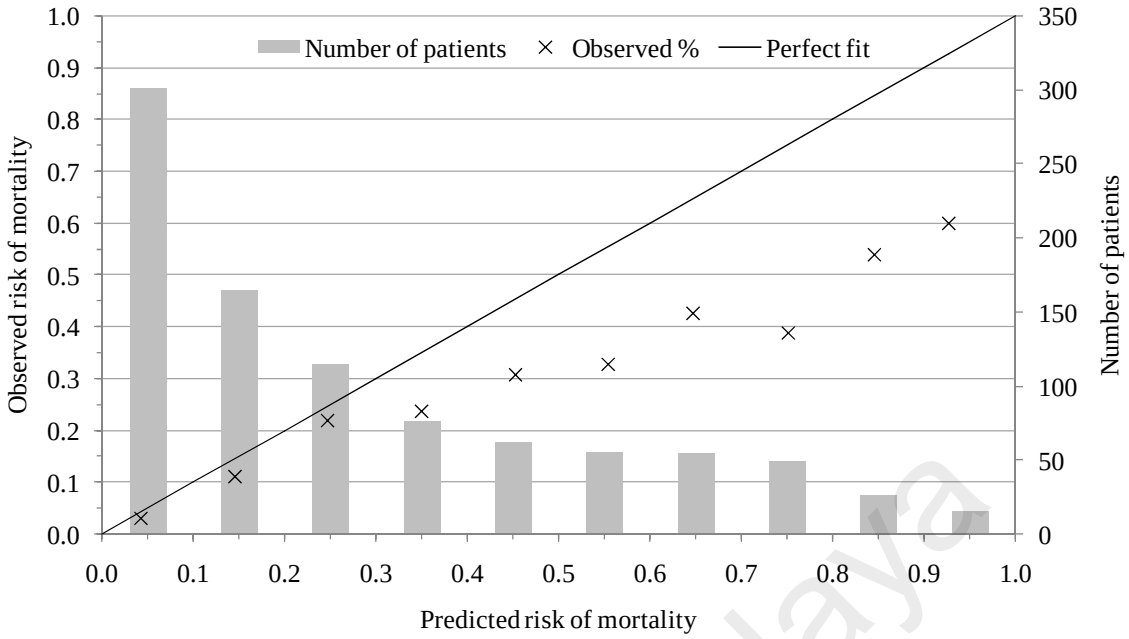


Figure 4.16: Calibration curve to compare observed and predicted in-ICU mortality rates across 10% intervals of predicted risk.

From the calibration curve, model fit is considered as perfect when the observed values lie exactly on the diagonal line. The calibration curve indicated that model fit appeared to be acceptable for the first three risk categories. However, predictions started to be inaccurate from the fourth risk category onwards, where the observed outcomes were much lower than predicted outcomes. The APACHE IV model appeared to overestimate in-ICU mortality, especially for mid to high risk ranges. A decreasing trend was observed on the number of patients as the predicted mortality risk increased. The majority of HSA ICU patients were in the first three groups and were associated with lower risks of death.

These findings suggested that APACHE IV was not suitable for application in HSA ICU. Despite having good discrimination, the model's calibration in HSA ICU was very poor. There were obvious differences in the baseline characteristics between HSA ICU and APACHE IV data sets. These differences could potentially be the factors that influenced the performance of APACHE IV in the Malaysian ICU. Application of

APACHE IV without further customisation would lead to inaccurate predictions. Thus, in this study, a first-level customisation strategy was applied to improve model calibration. Although there was no change in discrimination, a significant improvement was observed in the customised model's calibration (see Table 4.7). The overall fit for the customised model was found to be much improved, with a non-significant Hosmer-Lemeshow $\hat{C}$ statistic of 6.39 (p-value = 0.78).

Table 4.7: Performance of APACHE IV and first-level customised model in HSA ICU.

	SMR	AUC	Hosmer-Lemeshow $\hat{C}$ statistic
APACHE IV	0.67	0.78	113 (p-value <0.0001)
Customised model	1.00	0.78	6.39 (p-value = 0.78)

4.3 Proposed models using Bayesian Markov Chain Monte Carlo approach

4.3.1 Variable selection

In view of the extremely low percentage of HSA ICU patients with existing comorbidities, presence of chronic health (yes/no) was introduced as a variable to replace the seven individual chronic health categories defined in APACHE IV. Due to the high overall prevalence of diabetes in HSA ICU patients, diabetes (yes/no) was included as a potential variable. Patients were classified into one of nine individual admission diagnoses specified in APACHE IV. Trauma was chosen as the reference category due to a large number of younger patients being admitted to HSA ICU under this category. Age, APS and pre-ICU length of stay variables were treated as continuous variables, whereas the other variables were categorical in nature. A square-root transformation was applied on pre-ICU length of stay variable because of positive skew.

The Bayesian MCMC approach was used for analysis of univariate models for each candidate variable. The results generated by WinBUGS for each candidate variable are shown in Table 4.8. At the univariate level, a variable was considered as statistically

significant based on two criteria, i.e. if the *p*-value for the likelihood ratio test was less than 0.25 and the credible interval did not contain the value 'zero'.

Table 4.8: Log odds ratios of univariate tests for variables under consideration.

Variable	Posterior mean	SE	MC error	75% Credible Interval
Age (continuous, in years)	0.0049	0.0002	1.59E-05	(-0.001, 0.011)
APS (continuous, in points)	0.0332	0.0001	3.00E-05	(0.029, 0.037)
Ethnicity				
Chinese	0.1313	0.0071	2.89E-04	(-0.116, 0.379)
Indian	-0.0921	0.0102	3.33E-04	(-0.449, 0.265)
Others	0.0065	0.0109	3.49E-04	(-0.374, 0.387)
ICU admission source				
Another special care unit	-0.0417	0.0095	3.33E-04	(-0.374, 0.291)
Operating room	-0.0594	0.0064	2.76E-04	(-0.281, 0.162)
Gender (female)	-0.6305	0.0065	5.01E-04	(-0.856, -0.405)
Presence of chronic health (yes)	0.5016	0.0064	4.10E-04	(0.279, 0.724)
No GCS score due to patient being sedated/paralysed	0.8117	0.0066	7.10E-04	(0.583, 1.040)
Emergency surgery	0.1299	0.0061	2.61E-04	(-0.081, 0.341)
Mechanical ventilation (yes)	2.58	0.0185	0.004016	(1.938, 3.222)
Pre ICU length of stay (square root, in days)	-0.0056	0.0043	2.06E-04	(-0.154, 0.143)
Diabetes (yes)	0.5389	0.0070	5.04E-04	(0.296, 0.781)
Chronic health groups				
AIDS	1.593	0.0624	0.002232	(-0.581, 3.767)
Cancer	-25.84	0.6248	0.01566	(-47.587, -4.094)
Cirrhosis	0.7163	0.0339	0.001282	(-0.465, 1.897)
Hepatic failure	-24.2	0.6426	0.01615	(-46.568, -1.833)
Immunosuppression	-0.4463	0.0297	7.56E-04	(-1.481, 0.589)
Leukaemia/multiple myeloma	2.626	0.0524	0.002437	(0.803, 4.449)
Lymphoma	1.591	0.0625	0.002294	(-0.585, 3.767)
Disease groups				
Cardiovascular	0.0777	0.0085	4.97E-04	(-0.217, 0.373)
Gastrointestinal	-0.0182	0.0104	5.25E-04	(-0.382, 0.345)
Genitourinary	-2.56	0.0279	0.001402	(-3.532, -1.588)
Haematologic	2.359	0.0525	0.002369	(0.533, 4.185)
Metabolic/endocrine	-0.182	0.0194	6.65E-04	(-0.856, 0.492)
Musculoskeletal/skin	-25.75	0.6258	0.01579	(-47.531, -3.969)
Neurologic	-0.9518	0.0109	8.49E-04	(-1.332, -0.571)
Respiratory	-0.2765	0.0094	5.46E-04	(-0.602, 0.0491)

Note: *p*-values for likelihood ratio tests for all variables were < 0.25.

SE: standard error; MC: Monte Carlo; GCS: Glasgow Coma Scale.

Further examination of the results revealed that the following variables were found to be statistically significant at the univariate level: APS, gender, presence of chronic health, no GCS score due to patient being sedated/paralysed, mechanical ventilation and diabetes. An odds ratio of 0.53 for gender variable indicated that the odds of dying for female patients were 53% less than the odds of dying for male patients. Patients with at least one comorbidity had higher odds of dying in ICU, with an odds ratio of 1.65. Similarly, diabetic patients were observed to have higher odds of dying in ICU, with an odds ratio of 1.71. The odds of dying in ICU for patients without GCS score was 2.25 times higher than the odds of dying for those with GCS score. The odds of dying in ICU also increased for patients with higher APS values (odds ratio = 1.03) and those under mechanical ventilation (odds ratio = 13.2). Other variables such as age, ethnicity, ICU admission source, emergency surgery and pre ICU length of stay were found to be not significant due to their failure in meeting the statistical criteria.

4.3.2 Proposed Bayesian models

All of the variables that were found to be significant at the univariate level were collectively fitted into various combinations of multivariable models using the Bayesian MCMC approach in WinBUGS. These models were classified into five main types (Type W, Type M, Type P, Type A and Type F). Table 4.9 shows the list of variables in selected models of the five different types.

Type W models

The first type (Type W) comprised the main effects models with APS variable. These models were fitted with variables that were found to be significant at the univariate level (gender, presence of chronic health/diabetes, no GCS score due to patient being sedated or paralysed, mechanical ventilation and APS). Other variables such as ethnicity, ICU

admission source, emergency surgery and pre ICU length of stay were collectively tested in different combinations of multivariable models and were found to be not statistically significant at 5% level of significance. These variables were thus omitted and were not listed in Table 4.9.

Table 4.9: Variables for various combinations of multivariable models.

Model	Type W				Type M		Type P			Type A				Type F				
	Main effects with APS				APS. mean		Abnormal physiological (yes/no)			Abnormal physiological (%)				Factor analysis				
Variable	W1	W2	W3	W4	M1	M2	P1	P2	P3	A1	A2	A3	A4	F1	F2	F3	F4	F5
age	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
gender	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
APS	X	X	X	X														
mechanical ventilation	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
no GCS score	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X		
presence of chronic health	X				X		X			X		X		X	X		X	
diabetes		X	X			X		X			X		X			X		X
ICU admission diagnosis	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
trauma			X															
APS x trauma interaction			X															
APS.mean					X	X												
abnormal heart rate							X	X	X									
abnormal temperature							X	X	X									
abnormal WBC							X	X	X									
abnormal BUN							X	X	X									
abnormal sodium							X	X	X									
abnormal albumin							X	X	X									
abnormal bilirubin							X	X	X									
abnormal pH-$PaCO_2$							X	X	X									
physio.ooc										X	X							
%abnormal heart rate												X	X					
%abnormal temperature												X	X					
%abnormal sodium												X	X					
%abnormal bilirubin												X	X					
%abnormal PaO_2												X	X					
%abnormal pH-$PaCO_2$												X	X					
Factor 1														X				
Factor 2														X	X	X	X	X
Factor 3														X				
Factor 4														X	X	X	X	X
Factor 5														X				

Models W1 and W2 were almost similar except for one variable, i.e. presence of chronic health variable in W1 was replaced with diabetes in W2. In both models, patients were grouped into one of nine individual admission diagnoses, with trauma being the reference category. In model W3, patients were classified as either being in

trauma or non-trauma group. An interaction term between trauma and APS was also included in W3. Variables that were found to be not statistically significant in models W1/W2 were removed, resulting in model W4.

Type M models

Models M1 and M2 were almost equivalent to models W1 and W2 respectively, except for a difference in the calculation of the APS variable. The APS in models M1 and M2 were identified as APS.mean since it was calculated through combining scores that were assigned to the mean value of each physiological variable, instead of the worst value within the first day of ICU admission.

Type P models

Instead of using the APS variable, the worst values for each physiological variable in ICU Day 1 were dichotomously coded as being normal/abnormal (following APACHE IV definitions) in the third type of models (Type P). Table 4.10 shows the results of univariate tests for each of the abnormal worst physiological variables.

Three variables (abnormal mean blood pressure, abnormal total respiratory rate and abnormal haematocrit) were found to be not statistically significant at the univariate level, and were not included in Type P models. The rest of the abnormal physiological variables were collectively assessed for their statistical significance at the multivariate level, where those that were found to be statistically significant were finally included in the Type P models. The abnormal physiological variables that met inclusion criteria were abnormal heart rate, abnormal temperature, abnormal white blood cell count, abnormal blood urea nitrogen, abnormal sodium, abnormal albumin, abnormal bilirubin and abnormal pH-$PaCO_2$ relationship.

Table 4.10: Log odds ratios of univariate analyses for abnormal physiological variables.

Physiological variable	Posterior mean	SE	MC error	75% Credible Interval	Significant
Abnormal heart rate	1.504	0.010	0.0015	(1.139, 1.869)	Yes
Abnormal mean blood pressure	0.246	0.020	0.0047	(-0.454, 0.946)	No
Abnormal temperature	1.100	0.007	0.0010	(0.862, 1.338)	Yes
Abnormal total respiratory rate	-0.088	0.006	0.0003	(-0.295, 0.119)	No
Abnormal haematocrit	-0.242	0.010	0.0011	(-0.580, 0.096)	No
Abnormal white blood cell count	0.738	0.007	0.0007	(0.512, 0.964)	Yes
Abnormal creatinine	0.583	0.006	0.0005	(0.373, 0.793)	Yes
Abnormal total urine output	0.343	0.007	0.0006	(0.085, 0.601)	Yes
Abnormal blood urea nitrogen	1.123	0.007	0.0009	(0.888, 1.358)	Yes
Abnormal sodium	0.930	0.007	0.0008	(0.692, 1.167)	Yes
Abnormal albumin	1.301	0.006	0.0011	(1.078, 1.524)	Yes
Abnormal bilirubin	0.946	0.008	0.0009	(0.657, 1.235)	Yes
Abnormal glucose	0.791	0.006	0.0007	(0.579, 1.004)	Yes
Abnormal PaO_2	1.551	0.010	0.0015	(1.187, 1.915)	Yes
Abnormal pH-$PaCO_2$ relationship	1.668	0.012	0.0018	(1.261, 2.075)	Yes

Note: *p*-values for likelihood ratio tests for all physiological variables were < 0.25.

SE: standard error; MC: monte carlo

Models P1 and P2 were almost equivalent to models W1 and W2 respectively, in terms of the variables, except that the APS variable was substituted with dichotomous abnormal physiological variables. Variables that were found to be not statistically significant in models P1/P2 were removed, resulting in model P3.

Type A models

The fourth type of models (Type A) considered another alternative approach in using the frequency of abnormal physiological observations to substitute the use of APS as a severity of illness indicator. Models A1 and A2 consisted of variables from the main effects model W1 and W2 respectively, except for APS. A new variable known as

physio.ooc was included as a replacement for APS in these two models. This *physio.ooc* variable represented the overall percentage of abnormal physiological observations, where a higher percentage indicated a higher severity of illness for a specific patient. This variable was calculated as the ratio of the number of abnormal physiological variables over the total physiological variables within the first day of ICU admission.

Table 4.11 shows the results of univariate models for each physiological variable, in terms of the percentage of abnormal observations within the first day of ICU admission. At 25% level of significance, all of the variables were tested to be significant at the univariate level, except for haematocrit and albumin. These variables that were significant were then collectively included into a multivariable model, together with other significant variables from the main effects model (except for APS). At the multivariate level, the following variables were found to be significant at 5% level of significance: heart rate, temperature, sodium, bilirubin, PaO_2 and pH-$PaCO_2$ relationship. These variables were included into models A3 and A4, together with other significant variables from the main effects models.

Table 4.11: Log odds ratios for percentage of abnormal physiological variables.

Percentage of abnormal physiological variable	Posterior mean	SE	MC error	75% Credible Interval	Significant
Heart rate	1.750	0.255	0.002	(1.460, 2.040)	Yes
Mean blood pressure	0.395	0.327	0.001	(0.019, 0.771)	Yes
Temperature	14.550	3.710	0.013	(10.284, 18.817)	Yes
Total respiratory rate	-0.573	0.354	0.001	(-0.981, -0.166)	Yes
Haematocrit	0.479	1.480	0.003	(-1.223, 2.181)	No
White blood cell count	5.392	1.624	0.005	(3.524, 7.260)	Yes
Creatinine	4.909	1.374	0.005	(3.329, 6.489)	Yes
Total urine output	0.343	0.224	0.001	(0.085, 0.601)	Yes
Blood urea nitrogen	6.732	1.357	0.006	(5.171, 8.293)	Yes
Sodium	15.260	3.420	0.014	(11.327, 19.193)	Yes
Albumin	-0.767	8.326	0.008	(-10.342, 8.808)	No
Bilirubin	8.606	2.189	0.008	(6.089, 11.123)	Yes
Glucose	1.885	0.548	0.002	(1.255, 2.515)	Yes
PaO_2	5.594	0.721	0.005	(4.765, 6.423)	Yes
pH-$PaCO_2$ relationship	5.786	0.729	0.005	(4.947, 6.625)	Yes

Type F models

Factor analysis approach was explored in the modelling of the fifth type of models (Type F), where factor scores were calculated for the standardised worst values of the physiological variables within the first day of ICU admission. Table 4.12 illustrates the rotated component matrix that followed a Varimax with Kaiser Normalisation rotation method for the physiological variables. The results suggested the removal of bilirubin variable. Factor analysis was then performed on the remaining variables and the new rotated component matrix (after removal of bilirubin) is shown in Table 4.13.

Table 4.12: Rotated Component Matrix (initial).

Standardised worst physiological variable	Component				
	1	2	3	4	5
Zscore (heart rate)	-0.112	0.591	0.246	0.386	0.091
Zscore (mean blood pressure)	-0.046	-0.589	-0.036	0.110	-0.321
Zscore (temperature)	-0.032	0.144	0.574	-0.065	0.409
Zscore (total respiratory rate)	-0.042	0.222	0.676	0.171	-0.261
Zscore (haematocrit)	0.002	0.041	-0.076	0.055	0.736
Zscore (white blood cell count)	0.093	-0.157	0.204	0.661	0.187
Zscore (creatinine)	0.885	0.041	-0.024	0.136	-0.080
Zscore (urine output)	-0.610	0.070	-0.246	0.211	-0.247
Zscore (blood urea nitrogen)	0.891	0.177	-0.038	0.201	-0.068
Zscore (sodium)	-0.250	0.316	-0.626	0.133	0.097
Zscore (albumin)	-0.198	-0.703	0.113	0.072	0.187
Zscore (bilirubin)	0.000	0.274	0.200	-0.194	-0.241
Zscore (glucose)	0.079	0.060	-0.186	0.646	-0.103

Table 4.13: Rotated Component Matrix (bilirubin removed).

Standardised worst physiological variable	Component				
	1	2	3	4	5
Zscore (heart rate)	-	0.607	0.260	0.365	0.066
Zscore (mean blood pressure)	-	-0.604	-0.046	0.131	-0.290
Zscore (temperature)	-	0.153	0.576	-0.077	0.399
Zscore (total respiratory rate)	-	0.227	0.694	0.150	-0.281
Zscore (haematocrit)	0.004	0.041	-0.100	0.076	0.770
Zscore (white blood cell count)	0.094	-0.167	0.184	0.689	0.234
Zscore (creatinine)	0.886	0.031	-0.028	0.143	-0.071
Zscore (urine output)	-	0.058	-0.253	0.228	-0.224
Zscore (blood urea nitrogen)	0.891	0.173	-0.036	0.202	-0.068
Zscore (sodium)	-	0.326	-0.625	0.131	0.084
Zscore (albumin)	-	-0.707	0.096	0.092	0.219
Zscore (glucose)	0.078	0.073	-0.179	0.640	-0.119

Five factors were extracted based on the Principal Component Analysis approach, where variables (shaded in Table 4.13) were clustered into similar groups according to their loadings. Figure 4.17 depicts the physiological variables that were grouped into five factors. The component score coefficient matrix for all variables in Factors 1-5 is presented in Table 4.14. These values were derived using data from 916 admissions in year 2009.

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
❑ creatinine ❑ urine output ❑ blood urea nitrogen	❑ heart rate ❑ mean blood pressure ❑ albumin	❑ temperature ❑ total respiratory rate ❑ sodium	❑ white blood cell count ❑ glucose	❑ haematocrit

Figure 4.17: Five factors for worst values of physiological variables.

Table 4.14: Component Score Coefficient Matrix for all variables in five factors.

Standardised worst physiological variable	Component				
	1	2	3	4	5
Zscore (heart rate)	-0.117	0.386	0.168	0.277	0.026
Zscore (mean blood pressure)	-0.012	-0.413	0.008	0.157	-0.250
Zscore (temperature)	-0.040	0.081	0.385	-0.074	0.330
Zscore (total respiratory rate)	-0.083	0.128	0.505	0.122	-0.301
Zscore (haematocrit)	0.016	0.008	-0.113	0.053	0.706
Zscore (white blood cell count)	-0.0005	-0.177	0.124	0.587	0.203
Zscore (creatinine)	0.426	-0.020	-0.064	0.058	-0.048
Zscore (urine output)	-0.309	0.063	-0.136	0.232	-0.206
Zscore (blood urea nitrogen)	0.419	0.073	-0.076	0.098	-0.049
Zscore (sodium)	-0.109	0.246	-0.449	0.102	0.096
Zscore (albumin)	-0.079	-0.492	0.094	0.138	0.206
Zscore (glucose)	-0.002	0.013	-0.125	0.531	-0.105

These five factors were included into model F1, together with other variables (age, gender, mechanical ventilation, no GCS score, presence of chronic health and ICU admission diagnosis). However, results for model F1 using WinBUGS revealed that only Factor 2 and Factor 4 were statistically significant. Factor 1, Factor 3 and Factor 5 were observed to be not significant at 5% level of significance and were thus, excluded

in models F2, F3, F4 and F5. Factor 2 and Factor 4 were then included into various combinations of models F2, F3, F4 and F5. Models F2 and F4 have the same variables except for the difference in presence of chronic health (Model F2) and diabetes (Model F4). Models F3 and F5 are similar to models F2 and F4 respectively, with exception of one variable (absence of GCS score), which was removed since it was found to be not significant in both models F2 and F4. The component score coefficient matrix for the variables in Factor 2 and Factor 4 is shown in Table 4.15. The scores for Factor 2 and Factor 4 were calculated for each ICU patient as follows:-

Factor 2 score = -0.435*Zscore(heart rate) + 0.494*Zscore(mean blood pressure) + 0.535*Zscore(albumin) + 0.088*Zscore(white blood cell count) - 0.026*Zscore(glucose), (4.1)

Factor 4 score = 0.244*Zscore(heart rate) + 0.042*Zscore(mean blood pressure) + 0.156*Zscore(albumin) + 0.691*Zscore(white blood cell count) + 0.559*Zscore(glucose). (4.2)

Table 4.15: Component Score Coefficient Matrix for variables in Factor 2 and Factor 4.

Standardised worst physiological variable	Component	
	2	4
Zscore (heart rate)	-0.435	0.244
Zscore (mean blood pressure)	0.494	0.042
Zscore (albumin)	0.535	0.156
Zscore (white blood cell count)	0.088	0.691
Zscore (glucose)	-0.026	0.559

4.3.3 Performance and Validation Results of Proposed Models

The parameter estimates and odds ratios for the risk equations in type W, M, P, A and F models are shown in Appendix F (Tables F1-F4), Appendix G (Tables G1 and G2), Appendix H (Tables H1-H3), Appendix I (Tables I1-I4) and Appendix J (Tables J1-J5) respectively. Overall, these results suggested that age (10-year increments) had negligible effect on mortality risk, with odds ratios close to 1.0 in all the different types of models. Other factors such as increasing APS (10-unit increments), being male,

presence of chronic health, diabetes, being mechanically ventilated and absence of GCS information due to patient being sedated or paralysed, were found to have positive association with in-ICU mortality risk.

The APS variable was observed to be a significant and important predictor of in-ICU deaths in all of the type W models. The estimated odds ratio of APS for the type W models revealed that for an increase of ten points in APS, the odds of dying in HSA ICU increases by approximately 50%. Being mechanically ventilated, presence of chronic health and diabetes were found to be not significant at 5% level of significance in models W1 and W2 respectively. However, absence of GCS information was observed to be hugely significant in all type W models, with odds ratios of 5.75, 5.79, 6.70 and 6.49 in models W1-W4 respectively. The results also suggested that the odds of dying in ICU for female patients were about 50% lower than the odds of dying for males, with odds ratios approximately 0.5 for all type W models. In model W3, patients with trauma had a lower risk of dying compared to other patients. Positive interaction was observed between APS and trauma, where this interaction term was found to be significant in this model. Although interactions between trauma and other variables were tested, they were not statistically significant and therefore were not included in model W3.

The same set of variables that were observed to be significant in type W models was also found to be significant in type M models. The APS.mean variable was observed to be significant in models M1 and M2, where the estimated coefficients were comparable to the APS variable in type W models. There were also not much of differences in the estimated coefficients for other variables between type M and type W models.

The following variables were found to be significant at 5% level of significance in the type P models: gender, mechanical ventilation and abnormal dichotomous

physiological variables (heart rate, temperature, white blood cell count, blood urea nitrogen, sodium, albumin, bilirubin, ph-$PaCO_2$ relationship). Age and absence of GCS score information due to patient being sedated or paralysed were not significant in all of the type P models. In addition, presence of chronic health and diabetes was also not significant in models P1 and P2 respectively.

Gender, absence of GCS score and being mechanically ventilated were important predictors in type A models. Age and presence of chronic health/diabetes appeared to have negligible effect in these models. The physio.ooc variable in models A1 and A2 was observed to have strong positive association with risk of dying in the ICU, although the odds ratio was unnaturally large. In models A3 and A4, the following percentages of abnormal physiological variables were found to be significant at 5% level of significance: heart rate, temperature, sodium, bilirubin, PaO_2 and pH-$PaCO_2$ relationship. The estimates for temperature, sodium and bilirubin appeared to be extremely large, resulting in large odds ratios. These results were mainly due to a large percentage of input data being close to/ equal to the value of zero for these variables.

Variables that were found to be important in type F models included gender, being mechanically ventilated, presence of chronic health/diabetes. Age and absence of GCS score were not significant in all the type F models. Out of the five factors in models F1, only Factor 2 and Factor 4 were observed to be significant at 5% level of significance. Physiological variables that were grouped in Factor 2 were heart rate, mean blood pressure and albumin, whereas variables that were included in Factor 4 were white blood cell count and glucose. These two factors were subsequently verified as statistically significant in models F2-F5.

Table 4.16 shows the validation results and performance indicators of the models for Types W, M, P, A and F. Models W1 and W2 generally performed well, with good discrimination (AUC ≥ 0.8) and calibration (p-values > 0.05 in the Hosmer-

Lemeshow tests). The DIC values for both models were also almost equivalent, suggesting that there was no real improvement in model fit when "presence of chronic health" variable in model W1 was replaced with "diabetes" in model W2. However, model W1 slightly edged model W2 in the SMR and Brier Score measures, where the SMR in model W1 was closer to 1.0 and the Brier Score in model W1 was lower. Although removal of variables that were found to be not significant in model W1 and W2 resulted in a slightly lower DIC value in model W4, this difference was too small to be considered significant. The discrimination and calibration in model W4 was also comparable to models W1 and W2, although model W4 had the worst SMR, AUC and Brier Score among the three models.

Table 4.16: Performance indicators of the five different types of models.

Model		DIC	SMR (95% CI)	AUC (95% CI)	HL statistic (p-value)	Brier Score
W	W1	696.44	0.89 (0.61, 1.25)	0.810 (0.748, 0.862)	6.56 (p=0.58)	0.113
	W2	696.31	0.88 (0.61, 1.24)	0.808 (0.746, 0.861)	8.15 (p=0.42)	0.114
	W3	696.17	0.95 (0.65, 1.34)	0.792 (0.728, 0.846)	18.43 (p=0.0182)	0.113
	W4	696.00	0.87 (0.60, 1.22)	0.805 (0.742, 0.858)	8.19 (p=0.42)	0.115
M	M1	729.91	0.93 (0.64, 1.30)	0.801 (0.738, 0.855)	12.87 (p=0.12)	0.118
	M2	730.07	0.92 (0.63, 1.29)	0.798 (0.735, 0.852)	13.16 (p=0.11)	0.120
P	P1	720.47	0.94 (0.65, 1.32)	0.802 (0.740, 0.856)	6.85 (p=0.55)	0.119
	P2	720.33	0.94 (0.65, 1.32)	0.797 (0.734, 0.851)	8.13 (p=0.42)	0.122
	P3	723.85	0.99 (0.68, 1.39)	0.807 (0.744, 0.860)	4.75 (p=0.78)	0.118
A	A1	721.68	0.80 (0.55, 1.12)	0.835 (0.775, 0.884)	8.65 (p=0.37)	0.112
	A2	721.79	0.80 (0.55, 1.12)	0.833 (0.773, 0.883)	8.65 (p=0.37)	0.113
	A3	714.69	0.77 (0.53, 1.08)	0.810 (0.748, 0.863)	12.12 (p=0.15)	0.119
	A4	714.21	0.75 (0.52, 1.05)	0.793 (0.729, 0.847)	21.68 (p=0.0055)	0.126
F	F1	733.12	0.92 (0.63, 1.29)	0.763 (0.697, 0.821)	28.56 (p=0.0004)	0.127
	F2	734.09	0.91 (0.63, 1.28)	0.744 (0.676, 0.803)	20.19 (p=0.0096)	0.128
	F3	734.00	0.90 (0.62, 1.26)	0.739 (0.671, 0.799)	16.98 (p=0.03)	0.131
	F4	733.44	0.92 (0.64, 1.30)	0.741 (0.674, 0.801)	14.83 (p=0.0625)	0.128
	F5	732.98	0.92 (0.63, 1.28)	0.736 (0.668, 0.796)	18.43 (p=0.018)	0.131

On the other hand, model W3 had the worst discrimination (AUC < 0.8) and poor calibration across subgroups of patients with different risk profiles (p-value < 0.05

in the Hosmer and Lemeshow goodness-of-fit test). These findings suggested that it was still better to retain the original classification of ICU admission diagnoses in models W1, W2 and W4. There was no benefit in using a simplified classification of trauma/non-trauma to replace the ICU admission diagnoses in model W3.

The performances of the Type M models were inferior to the corresponding Type W models. This suggested that the use of APS.mean variable in models M1 and M2 did not improve the performance of the models, as compared to the use of APS variable. Although both models M1 and M2 registered slightly better SMR values, their DIC values and Brier Scores were considerably much higher than their counterparts (models W1 and W2). Both type M models exhibited similar performances in discrimination and calibration as their corresponding type W models, although model M1 appeared to slightly edge model M2 in the performance indicators.

Results that were obtained in Type P models suggested that replacing the APS variable with dichotomous abnormal physiological variables offered no substantial improvement in model performance. A comparison of the type P models revealed no noticeable differences in the performances of the three models. Model P1 appeared to be the best among the three type P models, with slightly better DIC, SMR, AUC and Brier Score values. However, the performance of model P3 was also found to be quite comparable to that of model P1. Model P3 can also be chosen as the best Type P model based on the criteria of parsimony.

A comparison between models A1 and W1 revealed no significant improvement in the overall model fit when the APS variable was replaced with a new physio.ooc variable. Although model A1 performed equally well in discrimination and calibration (p-value > 0.05 in the Hosmer-Lemeshow test and a low Brier score), its DIC and SMR values were not as good as that of model W1. Similarly, model A2 was also inferior to model W2, in terms of DIC and SMR values, despite exhibiting good discrimination

and calibration properties. The performance of models A3 and A4 (with individual percentages of abnormal physiological observations) was inferior to that of models P1 and P2 (with individual dichotomous abnormal physiological observations). Despite having better DIC values and equivalent discriminatory abilities, models A3 and A4 were not as well-calibrated as models P1 and P2. Although model A3 exhibited better discrimination and calibration than model A4, the SMR values for both models were less than 0.8, indicating that the overall predicted mortality risk was higher than the overall actual mortality risk.

The performances of the factor analysis models (type F) were the worst among the five different types of models. These type F models had the highest DIC values and lower AUC values compared to other model types. Poor calibration was also observed in almost all of the type F models, with *p*-values < 0.05 in the Hosmer Lemeshow tests and Brier Scores > 0.125. There appeared to be no remarkable differences between the performances of models F2 and F4, as well as, models F3 and F5. These results suggested that models F4 and F5 were preferable to models F2 and F3 as they contained lesser number of variables. The discrimination (AUC values) and calibration (Hosmer-Lemeshow tests) of model F4 was slightly better than model F5, although the DIC value for model F5 was slightly better than model F4.

Judging from the performances of the various models in Table 4.16, the overall prediction accuracy in W type of models were better than the other types of models. In particular, the DIC values in type W models were much lower compared to other types of models. Type W models also generally had less number of parameters compared to other model types (P, A, and F). Using both principle of parsimony and the DIC indicator as the criteria for model selection, the choice of best model was narrowed to one of the type W models. The overall performance of model W1 was considered the

best among all the type W models, as well as, other types of models. Further detailed analysis and statistical modelling of model W1 is elaborated in the following section.

4.4 Model W1

4.4.1 Variables in Model W1

Model W1 was fitted with the following seven variables: age, gender, APS, mechanical ventilation, presence of chronic health, absence of GCS score and ICU admission diagnoses. All of these variables were found to be statistically significant at the univariate level except for age and ICU admission diagnoses. Although these two variables were not statistically significant, they were still included in model W1 based on their clinical importance. The effect of variables in model W1 was examined by looking at several variations of model W1, with one or more variables removed. Table 4.17 shows the comparison of the DIC, deviance and values of likelihood ratio tests (G) between model W1 and other variations of model W1 (with certain variable(s) removed).

Table 4.17: Model fit comparison between model W1 and other variants of model W1.

Model	DIC	Deviance	G	*df*	*p*-value
W1 (reference model)	696.44	636.8			
With constant term only	887.097	844.6	207.8	14	<0.0001
W1 without age variable	695.294	638.2	1.4	1	0.2367
W1 without mechanical ventilation	696.543	640.4	3.6	1	0.0578
W1 without presence of chronic health	696.671	640.8	4.0	1	0.0455
W1 without gender	702.235	646.7	9.9	1	0.0017
W1 without no GCS score	735.987	674.6	37.8	1	<0.0001
W1 without APS	809.919	744.2	107.4	1	<0.0001
W1 without disease groups	700.797	650.0	13.2	8	0.1052
W1 without age and disease groups	699.572	650.5	13.7	9	0.1334
W1 without age, disease groups and mechanical ventilation	698.834	651.3	14.5	10	0.1514

The performance of model W1 was evidently much better than the model that was fitted with a constant term only. At 5% level of significance, the *p*-values obtained

in the likelihood ratio tests suggested the importance of including the APS, gender, no GCS score and presence of chronic health as predictors in model W1. Mechanical ventilation was also considered as an important variable in model W1 since the *p*-value of the likelihood ratio test for the model without mechanical ventilation was just slightly more than 0.05. On the other hand, the *p*-values of the likelihood ratio tests that compared model W1 with models that excluded the age variable and ICU admission diagnoses respectively were observed to exceed 0.05. This implied that there was no advantage in including the age variable or ICU admission diagnoses in model W1.

Age and ICU admission diagnoses were then removed from model W1. A comparison of the likelihood ratio test revealed that the reduced model was as good as the full model (model W1), as indicated by a *p*-value > 0.05. Likewise, the reduced model with three variables removed (age, ICU admission diagnoses and mechanical ventilation) was also found to be almost equivalent to model W1, with a large *p*-value > 0.05, based on the likelihood ratio test with ten degrees of freedom. Although inclusion of age, ICU admission diagnoses and mechanical ventilation variable were considered to offer no contribution to model W1 from a statistical perspective, these variables were considered important from the practical viewpoint and were then included in model W1.

4.4.2 Comparison between Bayesian and frequentist estimates in Model W1

A comparison of the estimated regression coefficients and standard errors obtained through the Bayesian and frequentist maximum likelihood estimation (MLE) methods for model W1 is shown in Table 4.18. Differences in the estimates for some of the admission diagnoses were due to small sample sizes in these disease categories. Generally, there were no substantial differences between the Bayesian and MLE estimates for most of the variables in the model. This was most likely due to the data set being sufficiently large enough, particularly for the MLE method. Large number of

iterations in the Bayesian models also played an essential role in ensuring convergence of the Markov chains to their equilibrium distributions. The standard errors that were obtained through the Bayesian approach were consistently much smaller compared to the frequentist (MLE) standard errors in model W1. Moreover, the deviance value produced by the Bayesian method was also much smaller compared to the deviance obtained using the MLE method. This indicated that model fit using the Bayesian approach was much better than the frequentist (MLE) method.

Table 4.18: Bayesian and frequentist (MLE) estimations in model W1.

Variable	Bayesian			MLE
	Coefficient (95% CI)	SE	OR (95% CI)	Coefficient ± SE
Age	-0.004 (-0.019, 0.010)	0.0002	1.00 (0.98, 1.01)	-0.004±0.007
Gender (female)	-0.638 (-1.130, -0.162)	0.008	0.53 (0.32, 0.85)	-0.582±0.224
APS	0.043 (0.034, 0.053)	0.0002	1.04 (1.03, 1.05)	0.039±0.004
No GCS score	1.753 (1.202, 2.331)	0.01	5.77 (3.33, 10.29)	1.589±0.254
Mechanical ventilation	0.811 (-0.349, 2.170)	0.021	2.25 (0.71, 8.76)	0.688±0.584
With chronic health	0.322 (-0.198, 0.841)	0.009	1.38 (0.82, 2.32)	0.295±0.241
Admission diagnoses				
Cardiovascular	0.009 (-0.671, 0.689)	0.011	1.01 (0.51, 1.99)	-0.021± 0.317
Respiratory	-0.265 (-0.962, 0.425)	0.012	0.77 (0.38, 1.53)	-0.271±0.325
Gastrointestinal	-0.208 (-1.023, 0.595)	0.014	0.81 (0.36, 1.81)	-0.214±0.376
Neurologic	-0.598 (-1.355, 0.131)	0.013	0.55 (0.26, 1.14)	-0.559±0.347
Metabolic/ endocrine	-0.285 (-1.654, 0.985)	0.022	0.75 (0.19, 2.68)	-0.231±0.600
Hematologic	2.526 (-0.234, 5.463)	0.048	12.50 (0.79, 236)	2.637±1.397
Genitourinary	-1.943 (-3.945, -0.380)	0.03	0.14 (0.02, 0.68)	-1.651±0.785
Musculoskeletal/ skin	-3.441(-7.760, -0.210)	0.064	0.03 (0.0004, 0.81)	-6.172±6.438
Deviance	636.80			670.56

4.4.3 MCMC Diagnostics of Model W1

The MCMC diagnostics of model W1 were assessed through trace, density, Brooks-Gelman-Rubin (BGR), autocorrelation and quantile plots. The trace plots did not show any specific trends or irregularities (see Figure 4.18). Convergence of the three parallel

chains with different initial values was observed in the BGR plots in Figure 4.19. The initial values in the parallel chains were chosen as starting points that were slightly dispersed relative to the posterior distribution. The quantile plots in Figure 4.20 also revealed that the quantiles for the variables reached stable equilibrium. The density plots in Figure 4.21 suggested that the estimated posterior distributions of the variables followed normal distributions, while the extra binomial variation variable followed a uniform distribution. High autocorrelations were observed for APS, age, presence of chronic health, absence of GCS score, disease categories (cardiovascular, gastrointestinal, neurologic, respiratory), mechanical ventilation and the intercept term in model W1 (see Figure 4.22).

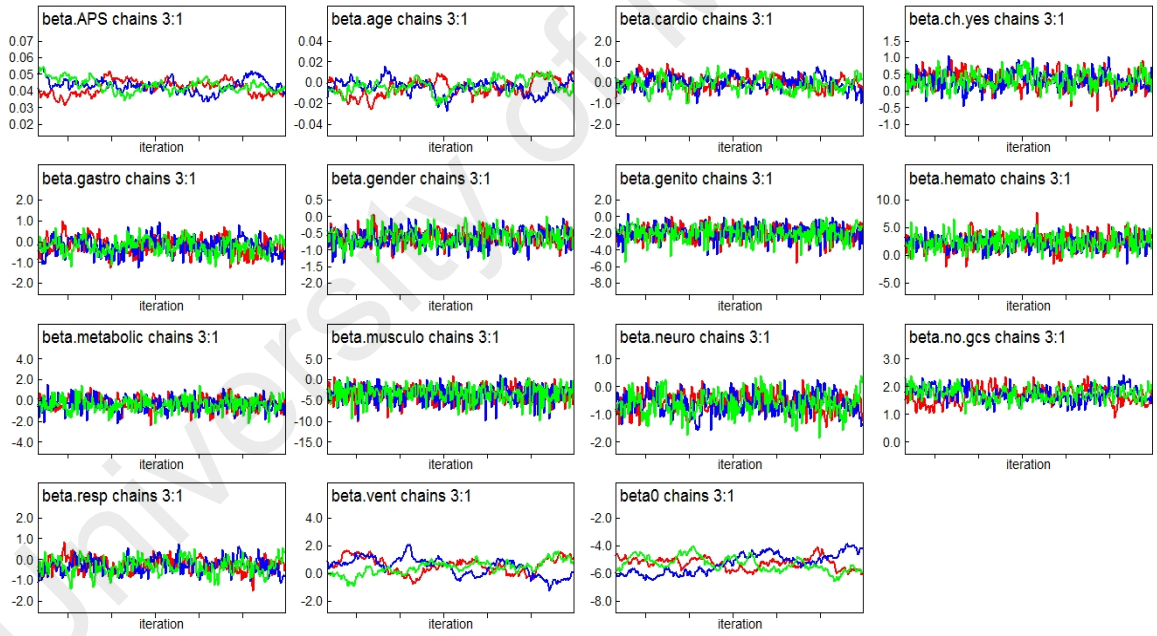


Figure 4.18: Trace plots for each variable in model W1.

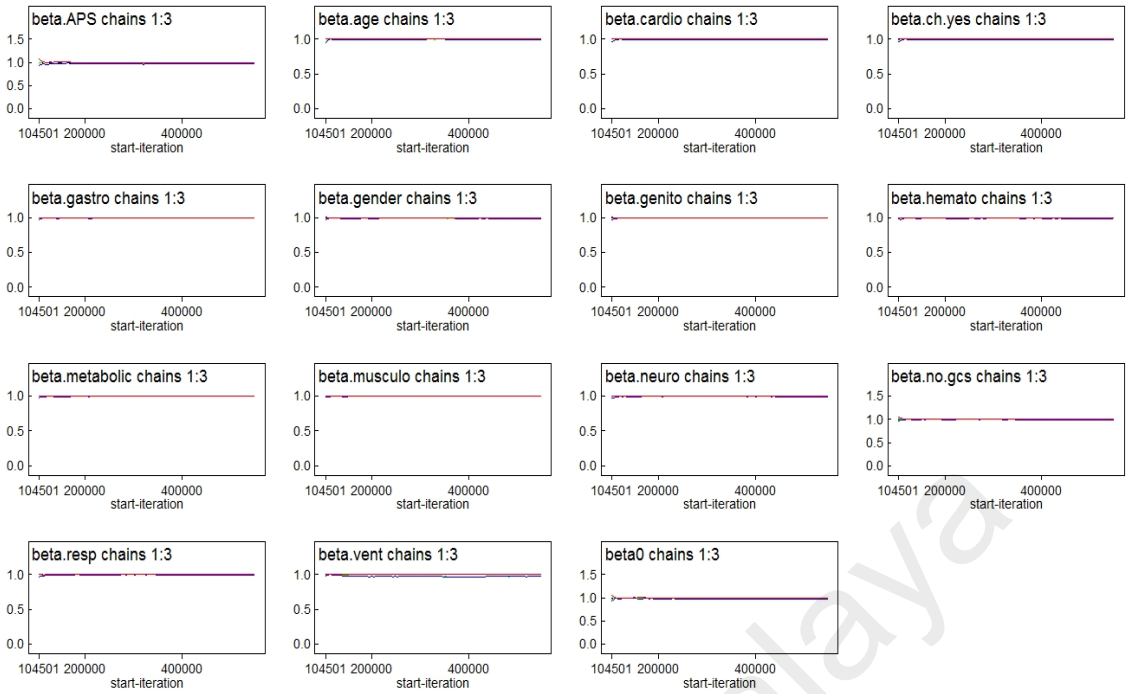


Figure 4.19: Brooks-Gelman-Rubin (BGR) plots for each variable in model W1.

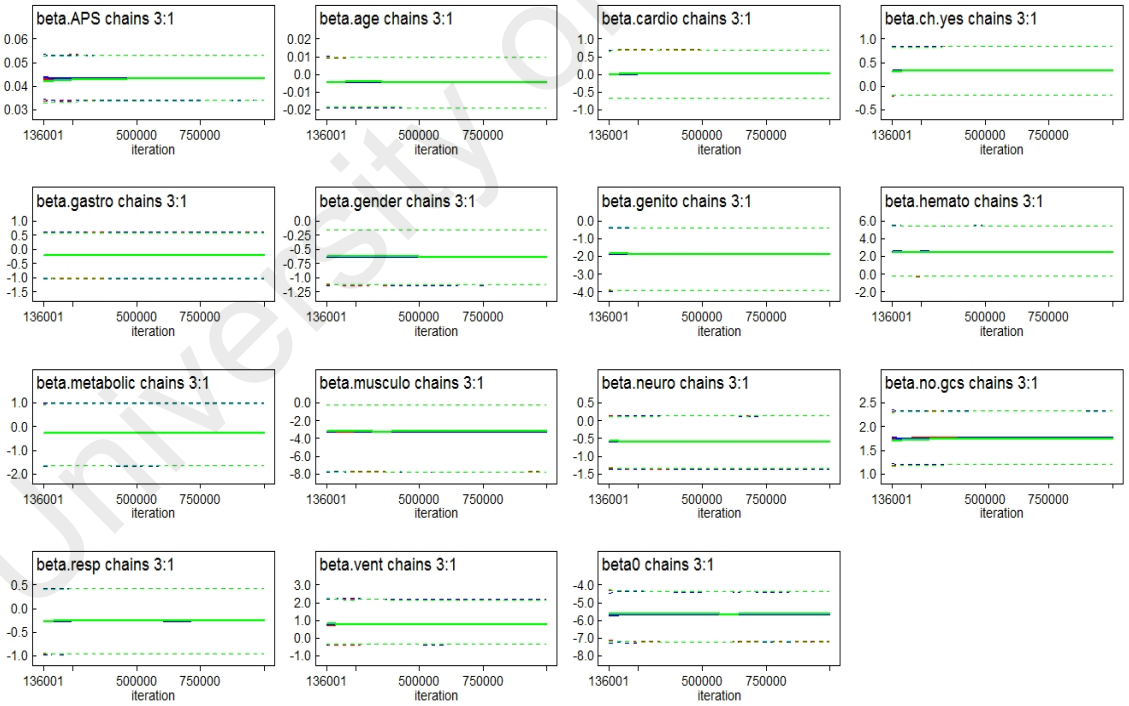


Figure 4.20: Quantile plots for each variable in model W1.

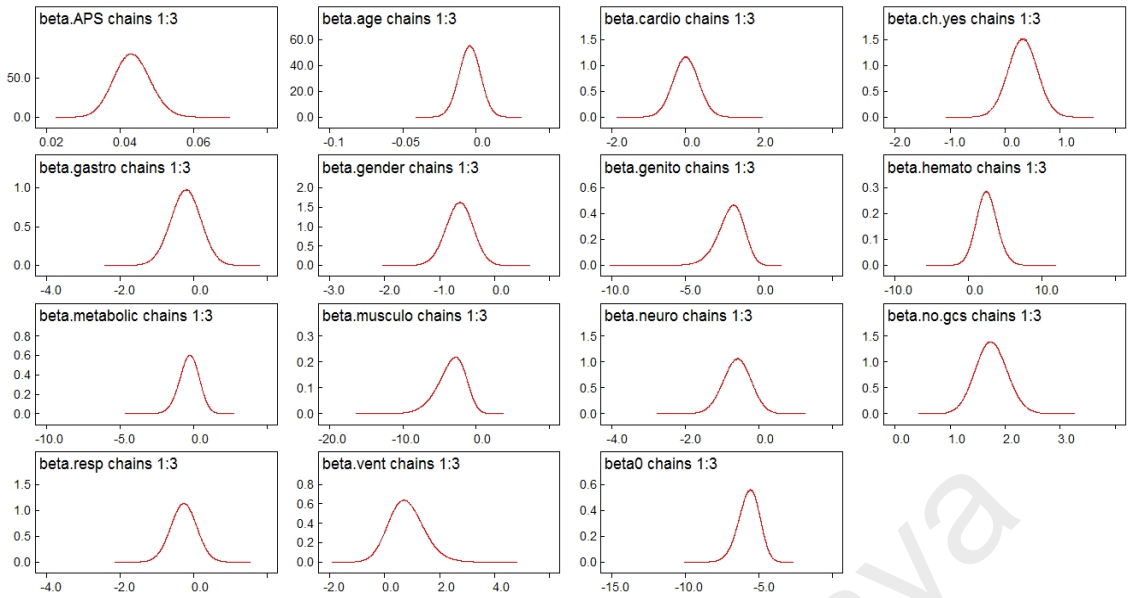


Figure 4.21: Density plots for each variable in model W1.

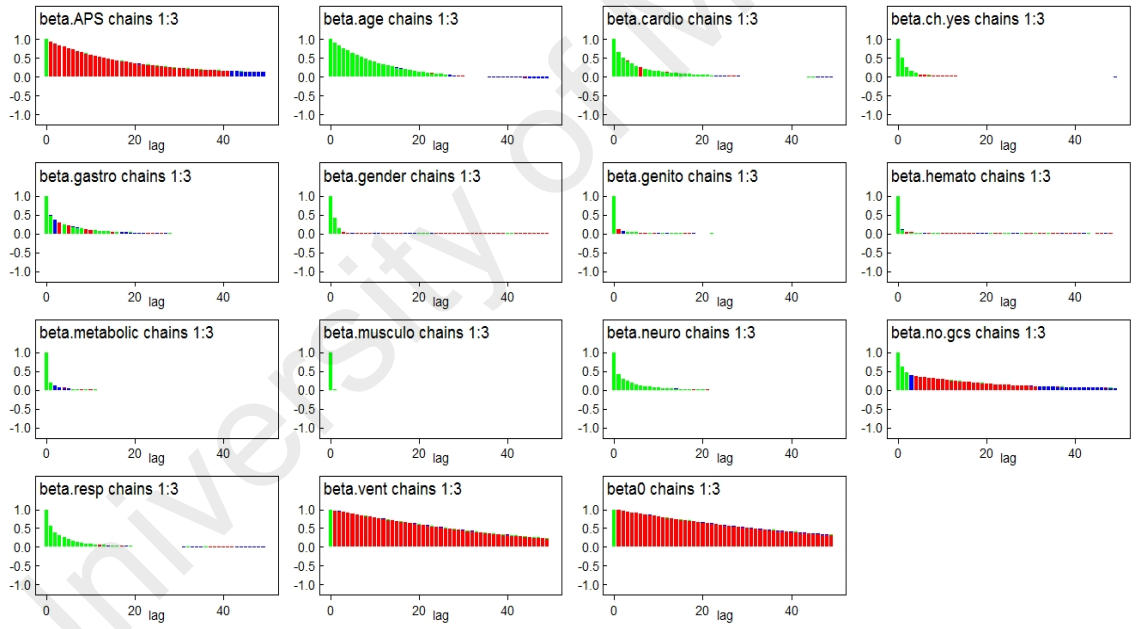


Figure 4.22: Autocorrelation plots for each variable in model W1.

The CODA results for model W1 are shown in Appendix K. The Gelman and Rubin Diagnostic test indicated model convergence with all variables having estimated potential scale reduction factors of 1.0 (Table K2). On the other hand, the Geweke output results (Table K1) revealed differences in the means of the first and last groups of iterations for a few variables in the third chain, where not all variables had Z values

within -2 and 2. Some variables also failed the Heidelberger and Welch diagnostic test (Table K4). Generated values from the Markov chains are considered to be independent when the dependence factor in the Raftery-Lewis test is equal to one (Ntzoufras, 2009). However, high values of dependence factor were observed for some variables in the Raftery-Lewis test (Table K3). These results suggested the possibility of high auto-correlations for these variables in model W1.

The thinning interval in model W1 was then adjusted in order to obtain more independent and less correlated samples. The minimum required number of burn-in samples and thinning interval were estimated by looking at the highest value of dependence factor in the Raftery-Lewis output in Table K3. Model W1 was then re-run with 200 burn-in samples and 50,000 subsequent iterations, with the thinning interval set as 60. The trace plots in Figure 4.23 and four convergence diagnostic tests suggested convergence for model W1 when the thinning interval was adjusted to 60 (see Tables K5-K8).

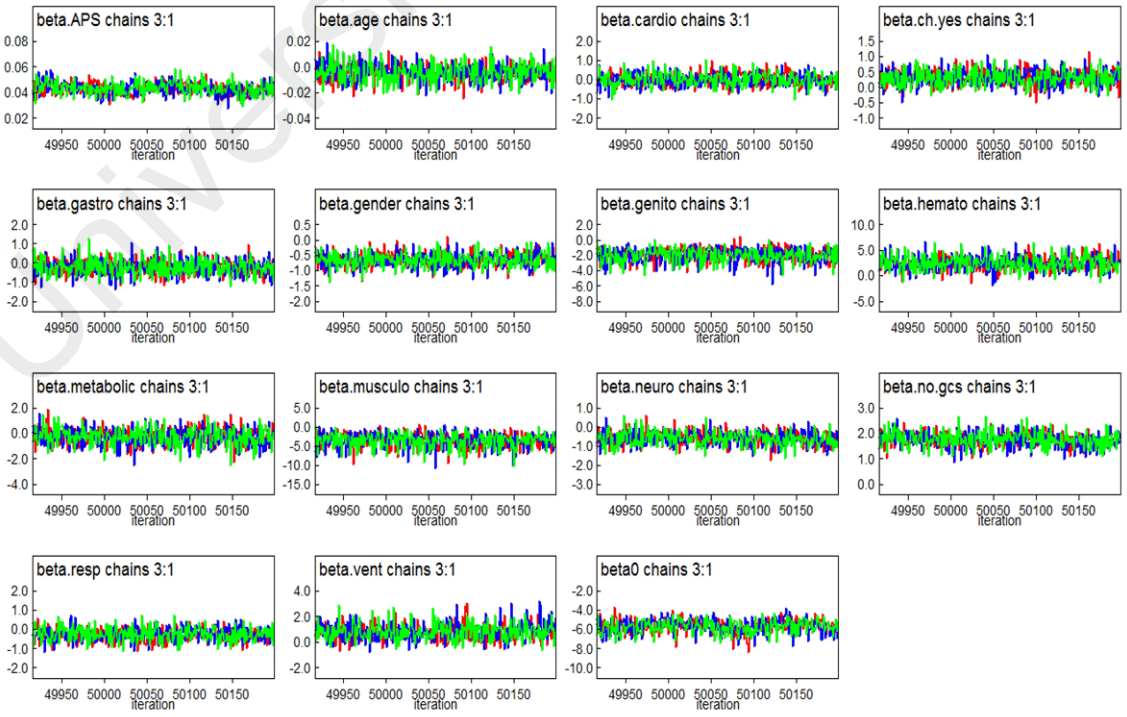


Figure 4.23: Trace plots in model W1 (thinning interval = 60).

The auto-correlations for the problematic variables were also resolved with the adjusted thinning interval (see Figure 4.24). However, there were no noticeable differences in the estimated posterior means and deviances for model W1 with thinning intervals 1 and 60 (see Table 4.19). This suggested that there were no differences in the performances of model W1 when the thinning interval was adjusted from 1 to 60.

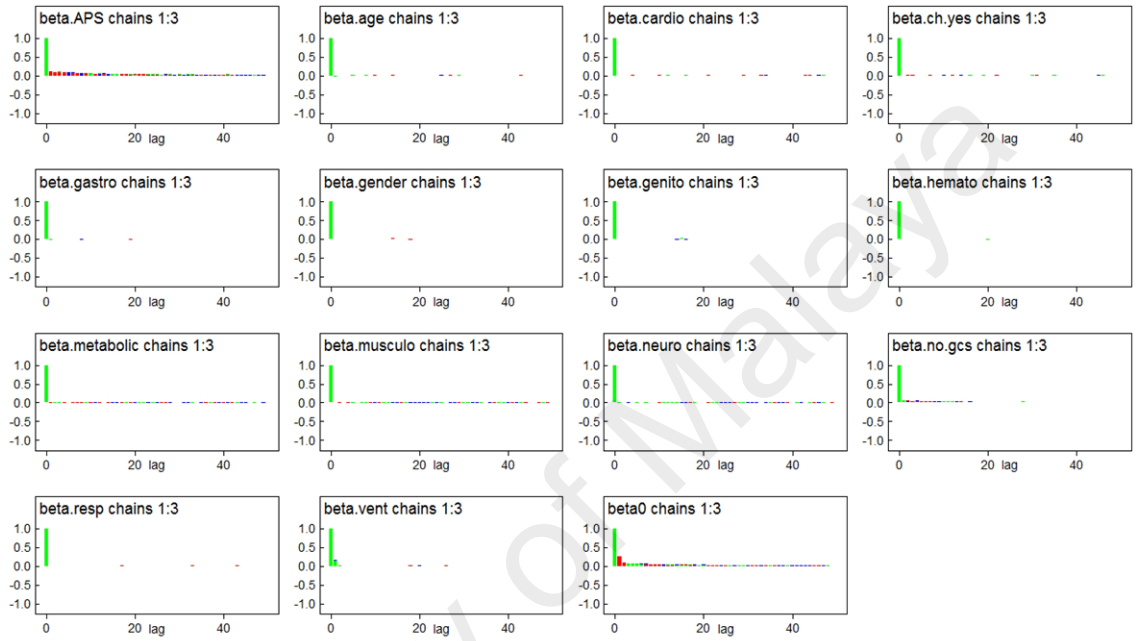


Figure 4.24: Autocorrelation plots in model W1 (thinning interval=60).

Table 4.19: Estimated posterior means of model W1 with thinning intervals 1 and 60.

Thinning interval / Variable	1	60
Age	-0.004	-0.004
Gender (female)	-0.638	-0.638
APS	0.043	0.043
Mechanical ventilation	0.811	0.804
No GCS score	1.753	1.749
With chronic health	0.322	0.323
ICU admission diagnoses		
Cardiovascular	0.009	0.008
Respiratory	-0.265	-0.265
Gastrointestinal	-0.208	-0.208
Neurologic	-0.598	-0.597
Metabolic/endocrine	-0.285	-0.285
Haematologic	2.526	2.519
Genitourinary	-1.943	-1.941
Musculoskeletal/skin	-3.441	-3.436
Deviance	636.8	638.1

4.4.4 Tests of Linearity for continuous variables in Model W1

Although age was found to be not significant, this variable was included due to its clinical relevance in all the multivariable models. Linearity checking was performed on age and APS since these were continuous variables in model W1. Figures 4.25 and 4.26 show the scatter plots of the logit term of model W1 versus age and APS respectively. To ascertain linearity of the plots, they were also fitted with a linear line and a smoothed LOWESS logit line. Both scatter plots do not appear to show patterns that indicate non-linearity. The LOWESS smoothed logit line in Figure 4.25 falls within the 95% confidence intervals of the fitted linear line, indicating a linear relationship between age and the logit term of model W1. The plot of standardised residuals across the standardised predictions in Figure 4.27 also does not reveal any patterns of non-linearity for age variable in model W1.

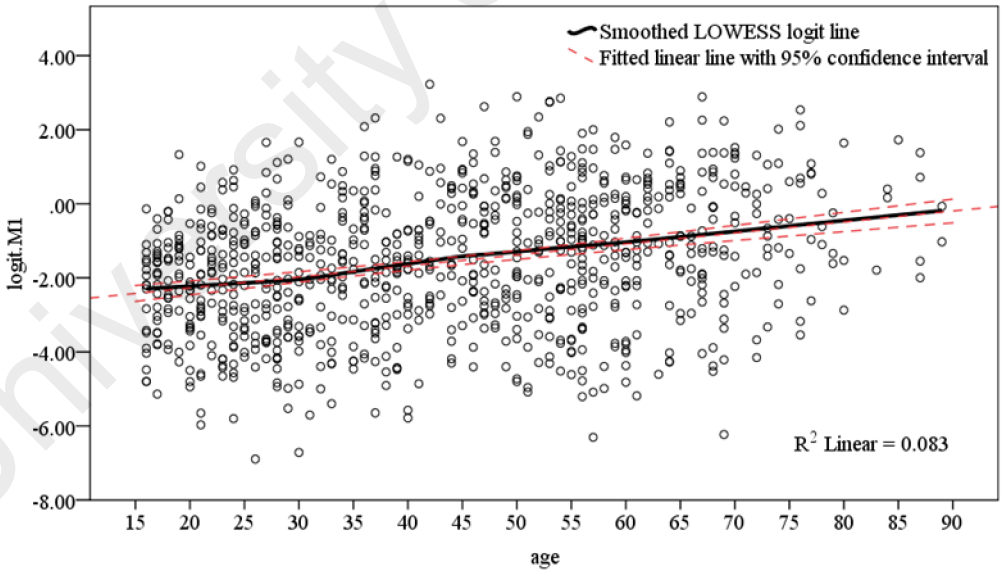


Figure 4.25: Plot of logit (model W1) versus age.

Figure 4.26 shows that the smoothed LOWESS logit line slightly falls beyond the 95% estimated confidence intervals of the linear line in the beginning and middle portions of the plot for APS variable in model W1. The pattern observed in the

corresponding residual plot in Figure 4.28 indicate the presence of some outliers (more than three standard deviations from the mean), which may have affected the behaviour of the plot in Figure 4.26.

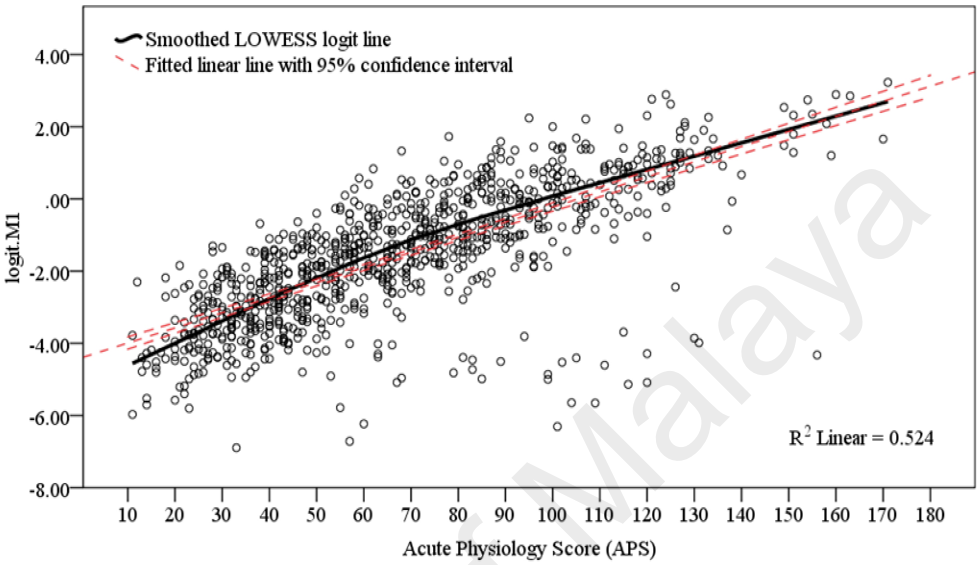


Figure 4.26: Plot of logit (model W1) versus APS.

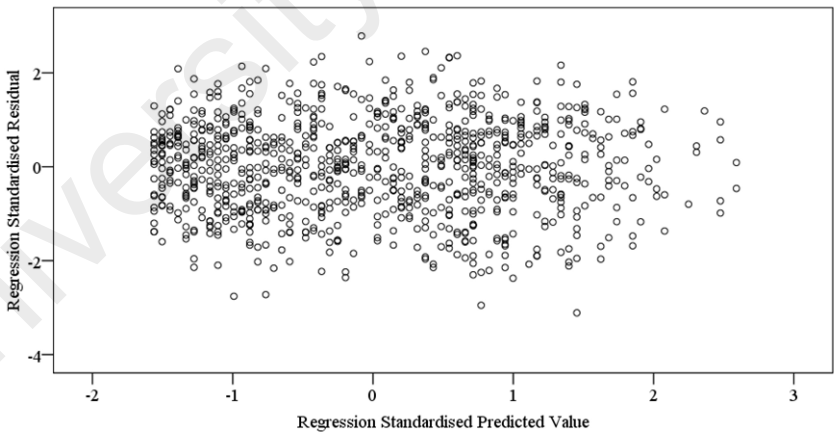


Figure 4.27: Plot of standardised residuals against standardised predictions for age variable in model W1.

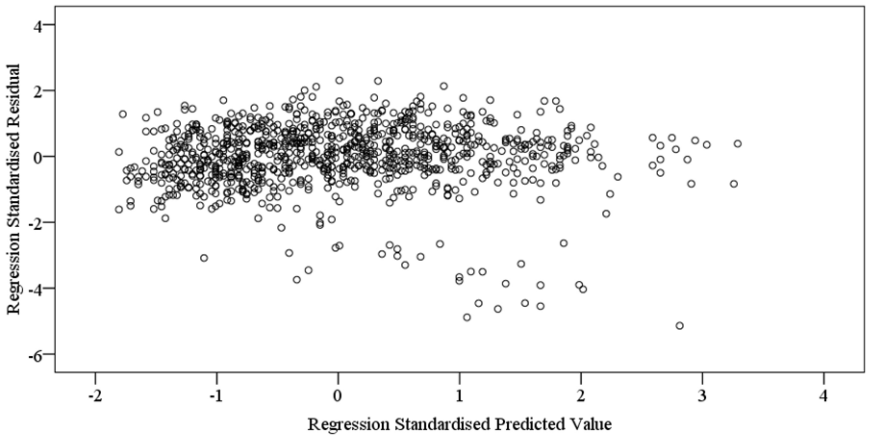


Figure 4.28: Plot of standardised residuals against standardised predictions for APS variable in model W1.

Further verification of the linearity assumption of APS in model W1 was performed using the method of design variables that was proposed in Hosmer and Lemeshow (2000). The APS variable in model W1 was substituted with a categorical variable that was based on four levels, using the quartiles as the cut points and the lowest quartile as the reference category. Table 4.20 shows the results of the quartile analyses of APS in model W1, whereas Figure 4.29 depicts the plot of estimated regression coefficients for quartile midpoints of APS in model W1. The 95% confidence intervals for the estimated regression coefficients for all quartiles did not contain the value of zero, indicating statistical significance of the APS variable. An increasing trend in the estimated regression coefficients in Figure 4.29 supported the linearity assumption for APS variable in model W1.

Table 4.20: Results of the quartile analyses of APS in model W1.

Quartile	1	2	3	4
Midpoint of APS	26.5	54	77.5	130.5
Estimated coefficient	0	1.005	1.93	3.711
95% confidence interval of regression coefficient		(0.1092, 1.971)	(1.026, 2.919)	(2.749, 4.77)

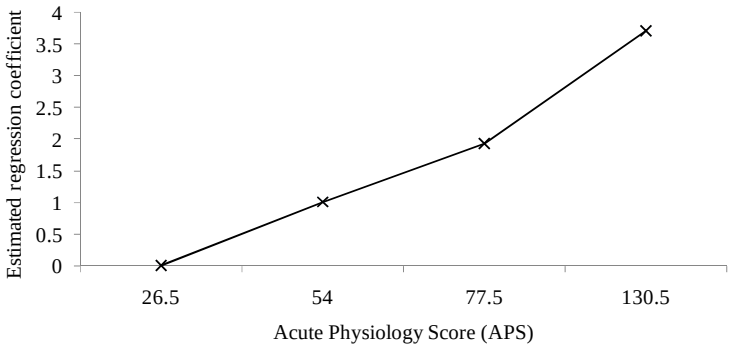


Figure 4.29: Plot of estimated coefficients for APS quartile midpoints in model W1.

Using a similar approach, the results of quartile analyses on age in model W1 is displayed in Table 4.21 and the corresponding plot of estimated coefficients is shown in Figure 4.30. There was no conclusive evidence to support or disprove the linearity assumption, as the plot indicated an initial decrease in the log odds, followed by an increase in the fourth coefficient. However, the 95% confidence intervals of the second, third and fourth quartile were overlapped and contained the value zero. These results supported the finding that age was not significant in model W1.

Table 4.21: Results of the quartile analyses of age in model W1.

Quartile	1	2	3	4
Midpoint of age	22	36	50.5	73.5
Estimated coefficient	0	-0.1931	-0.3001	-0.2119
95% confidence interval of regression coefficient		(-0.836, 0.444)	(-0.937, 0.329)	(-0.897, 0.469)

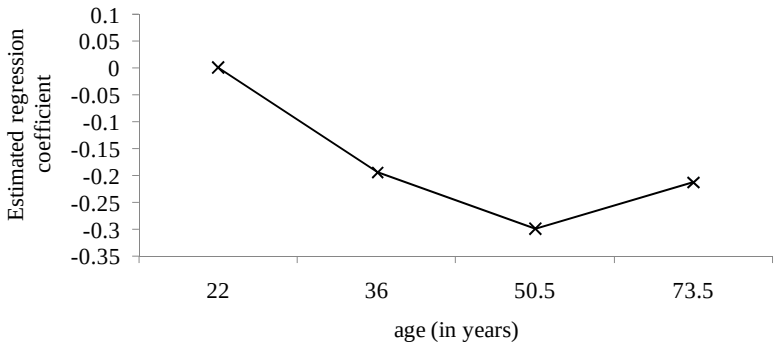


Figure 4.30: Plot of estimated coefficients for age quartile midpoints in model W1.

Linearity of the continuous variables in model W1 was also tested by including an additional non-linear term for age and APS in the logistic regression model. This was achieved by introducing a squared term for age and APS, i.e. age squared and APS squared, in addition to the original variables in model W1. Box-Tidwell transformation approach was also employed to the age and APS variables, resulting in the inclusion of new terms age × log(age) and APS × log(APS) in model W1. These models were then separately compiled and tested in WinBUGS. Table 4.22 displays the results of the models with the additional non-linear terms. Age was not significant in the models with additional non-linear terms, whereas APS was significant. For both W1 models with the additional non-linear terms, age^2 and APS^2, as well as, age × log(age) and APS × log(APS) were all found to be not significant at 5% level of significance. These findings supported the linearity assumption of age and APS variables in model W1.

Table 4.22: Parameter estimates of the additional non-linear terms in model W1.

Model type	Variable	Posterior Mean (95% Credible Interval)	Standard Error	Monte Carlo Error
W1 with squared terms	age	-0.0401 (-0.1040, 0.0180)	0.0010	0.0010
	age squared	0.0004 (-0.0002, 0.0011)	1.1×10^{-5}	1.4×10^{-5}
	APS	0.0638 (0.0294, 0.0995)	0.0006	0.001
	APS squared	-0.0001 (-0.0003, 0.0001)	3.1×10^{-6}	4.1×10^{-6}
W1 with log terms	age	-0.1557 (-0.3960, 0.1190)	0.0044	0.0074
	age × log(age)	0.0318 (-0.0255, 0.0082)	0.0009	0.0016
	APS	0.1103 (0.0052, 0.2522)	0.0021	0.0035
	APS × log(APS)	-0.0122 (-0.0385, 0.0071)	0.0004	0.0006

4.4.5 Tests of Interaction Effects in Model W1

Evaluation of the significance of interaction effects was conducted by separately fitting model W1 with each interaction term in WinBUGS, based on three multiple chains involving 1,000,000 iterations. Table 4.23 shows the combinations of interaction terms that were introduced in model W1 in order to test for presence of interaction between

variables. The estimates for the interaction terms and deviances for model W1 with the additional interaction terms are shown in Table 4.24. Details of changes in the estimates of the interaction terms are available in Appendix L. The likelihood ratio test statistic was computed by taking the difference between deviance values for model W1 (main effects model without interaction effect) and model W1 (with interaction effect).

Table 4.23: Plausible interactions between variables in model W1.

Variables	age	APS	gender	ch.yes	vent
age					
APS	×				
Gender	×	×			
presence of chronic health (ch.yes)	×	×	×		
mechanical ventilation (vent)	×	×	×	×	
absence of GCS score (no.gcs)	×	×	×	×	×

Table 4.24: Estimated regression coefficients of the interaction terms in model W1 and their corresponding deviance and likelihood ratio test statistics (G).

Model W1 with	Estimated coefficient (95% Credible Interval)	Deviance	G	*p*-value	Sig.
No interaction		636.8			
age×APS	-0.0001 (-0.0005, 0.0004)	638.8	-2.0	0.1573	No
age×gender	-0.0270 (-0.0550, 0.0003)	632.0	4.8	0.0285	No
age×chronic health	0.0234 (-0.0098, 0.0577)	645.4	-1.4	0.2367	No
age×ventilation	-0.0519 (-0.1103, 0.0054)	634.6	2.2	0.1380	No
age×absence of GCS	-0.0136 (-0.0393, 0.0122)	635.1	1.7	0.1923	No
APS×gender	-0.0052 (-0.0200, 0.0099)	637.7	-0.9	0.3428	No
APS×chronic health	-0.0044 (-0.0212, 0.0127)	636.5	0.3	0.5839	No
APS×ventilation	-0.0060 (-0.0541, 0.0440)	637.4	-0.6	0.4386	No
APS×absence of GCS	0.0024 (-0.0171, 0.0226)	639.3	-2.5	0.1138	No
gender×chronic health	0.0462 (-0.8342, 0.9221)	637.7	-0.9	0.3428	No
gender×ventilation	0.7137 (-0.7925, 2.2840)	636.0	0.8	0.3711	No
gender×absence of GCS	0.0476 (-0.8221, 0.9098)	639.9	-3.1	0.0783	No
ch.yes×ventilation	-1.1330 (-2.6080, 0.3208)	633.8	3.0	0.0833	No
ch.yes×absence of GCS	-0.0461 (-0.9239, 0.8318)	635.6	1.2	0.2733	No
vent×absence of GCS	0.1658 (-1.7060, 2.0280)	638.7	-1.9	0.1681	No

The significance of the interaction terms were evaluated based on the *p*-values of the likelihood ratio test and 95% posterior mean credible intervals for the estimated coefficients of the interaction terms. The deviance values of the models with additional

interaction terms were compared to deviance of the original model W1 (without any interaction term). Although slight improvements in deviance values were observed when age×gender, age×ventilation, age×absence of GCS, APS×chronic health, gender× ventilation, chronic health×ventilation and chronic health×absence of GCS interaction terms were individually included into model W1, differences in the deviance values were considered marginal and the interaction terms were found to be not significant based on their 95% credible intervals. All of the interaction terms were found to be not significant and there was no additional benefit in including interaction terms in model W1. Thus, the main effects model W1 with seven variables was considered as the final model proposed for application in HSA ICU.

4.4.6 Validation results and performance of Model W1

Model W1 was validated using data set from HSA ICU admissions between 1 January 2009 and 30 June 2009. The validation results and performance indicators of model W1 are summarised in Table 4.25. Model W1 achieved an SMR value of 0.887 (95% CI: 0.610, 1.245). This value indicated that the overall actual mortality was slightly lesser than the predicted mortality, although it was still quite close to one. The small Brier score in model W1 also indicated good overall accuracy. Figure 4.31 illustrates the receiver operating characteristic curve for model W1. Discrimination in model W1 was fairly good, with an AUC of 0.810 (95% CI: 0.748, 0.862) (see Table 4.26).

Table 4.25: Performance indicators of model W1 based on validation data set (n=195).

Performance indicator	Model W1
SMR (95% confidence interval)	0.887 (0.610, 1.245)
AUC (95% confidence interval)	0.810 (0.748, 0.862)
Brier Score	0.113
Hosmer-Lemeshow statistic (p-value)	6.56 (p-value = 0.5848)
Deviance Information Criterion (DIC)	696.44

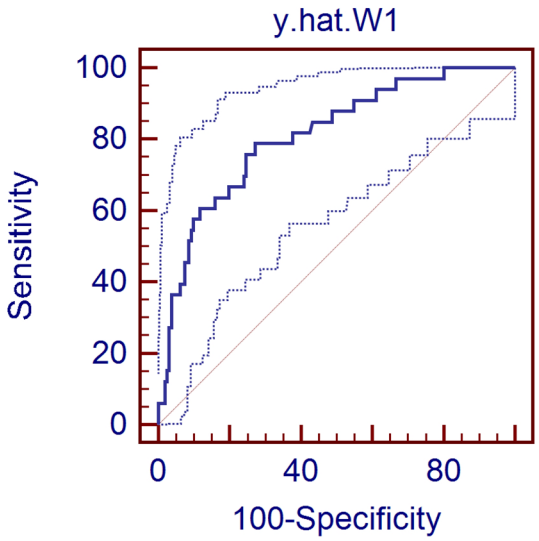


Figure 4.31: Receiver Operating Characteristic (ROC) curve for model W1.

Table 4.26: Area under receiver operating characteristic curve (AUC) for model W1.

Variable	y.hat.W1
Classification variable	actual
Sample size	195
Positive group: actual = 1	33
Negative group: actual = 0	162
Disease prevalence (%)	Unknown
Area under the ROC curve (AUC)	0.810
Standard Error[a]	0.0474
95% Confidence Interval[b]	0.748 to 0.862
z statistic	6.531
Significance level P (Area=0.5)	<0.0001

[a] Hanley & McNeil, 1982
[b] Binomial exact

A comparison of the actual number of deaths and model W1's predicted mortality rates for ten groups of patients with different risk groups is shown in Table 4.27. The risk groups were sorted in increasing order from the lowest (group 1) to the highest (group 10). Figure 4.32 shows the calibration curve of model W1 across the ten groups of HSA ICU patients. A closer inspection of the table and plot revealed close agreement between observed and predicted risk mortality for patients across most of the risk groups, except the eighth group. Model W1 slightly underpredicted mortality risk

for patients in the lower risk groups (second, third and fourth) and slightly overestimated mortality risk for mid- to high- risk patients.

Table 4.27: Observed and predicted (model W1) mortality rates across different risk categories within HSA ICU validation data set (n=195).

Risk category	Observed deaths Number (%)	Predicted deaths Number (%)	Difference (%)	Total patients
1	0 (0.0)	0.06 (0.3)	-0.3	20
2	1 (5.3)	0.2 (1.2)	4.0	19
3	1 (5.0)	0.6 (3.1)	1.9	20
4	2 (10.0)	1.2 (6.0)	4.0	20
5	1 (5.3)	1.9 (9.8)	-4.5	19
6	2 (10.0)	2.8 (14.0)	-4.0	20
7	4 (20.0)	4.0 (20.0)	0.0	20
8	3 (15.8)	5.9 (31.2)	-15.4	19
9	7 (36.8)	8.2 (43.0)	-6.2	19
10	12 (63.2)	12.3 (64.9)	-1.8	19

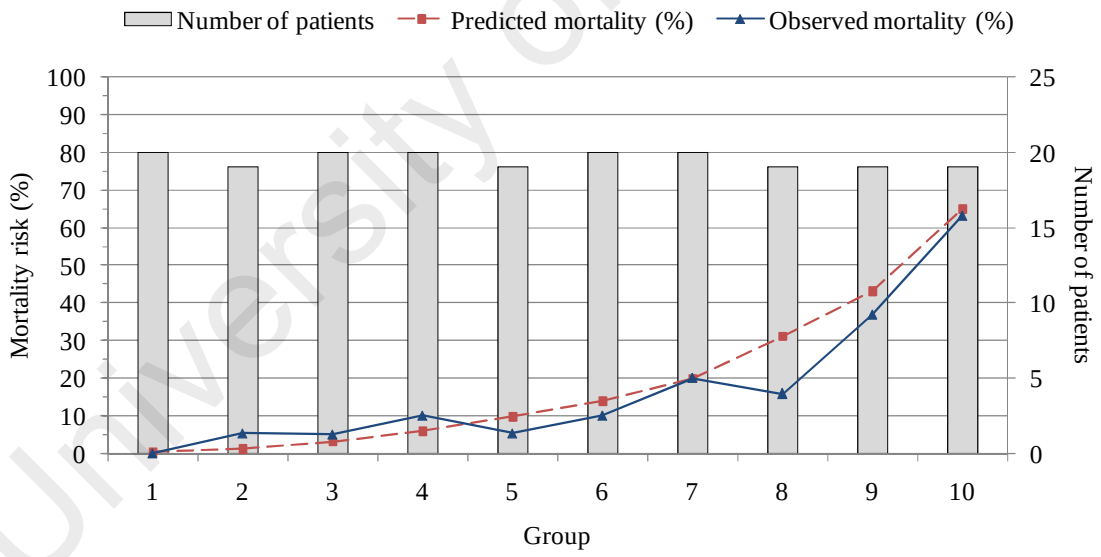


Figure 4.32: Calibration curve of model W1 based on validation data set (n=195).

4.5 Mortality Prediction using Logistic Regression Decision Tree Approach

All variables in model W1 were used in the construction of the decision tree, with the outcome variable being risk of death (mortality outcome upon discharge from ICU).

Figure 4.33 illustrates the decision tree that was derived from a training set of 916 patients. Development of the decision tree based on the CHAID algorithm generated a tree depth of four levels, with fifteen decision rules and nine terminal nodes. Age, gender, APS and absence of GCS were finally included in the decision tree model, whereas presence of chronic health, mechanical ventilation status and ICU admission diagnoses were omitted.

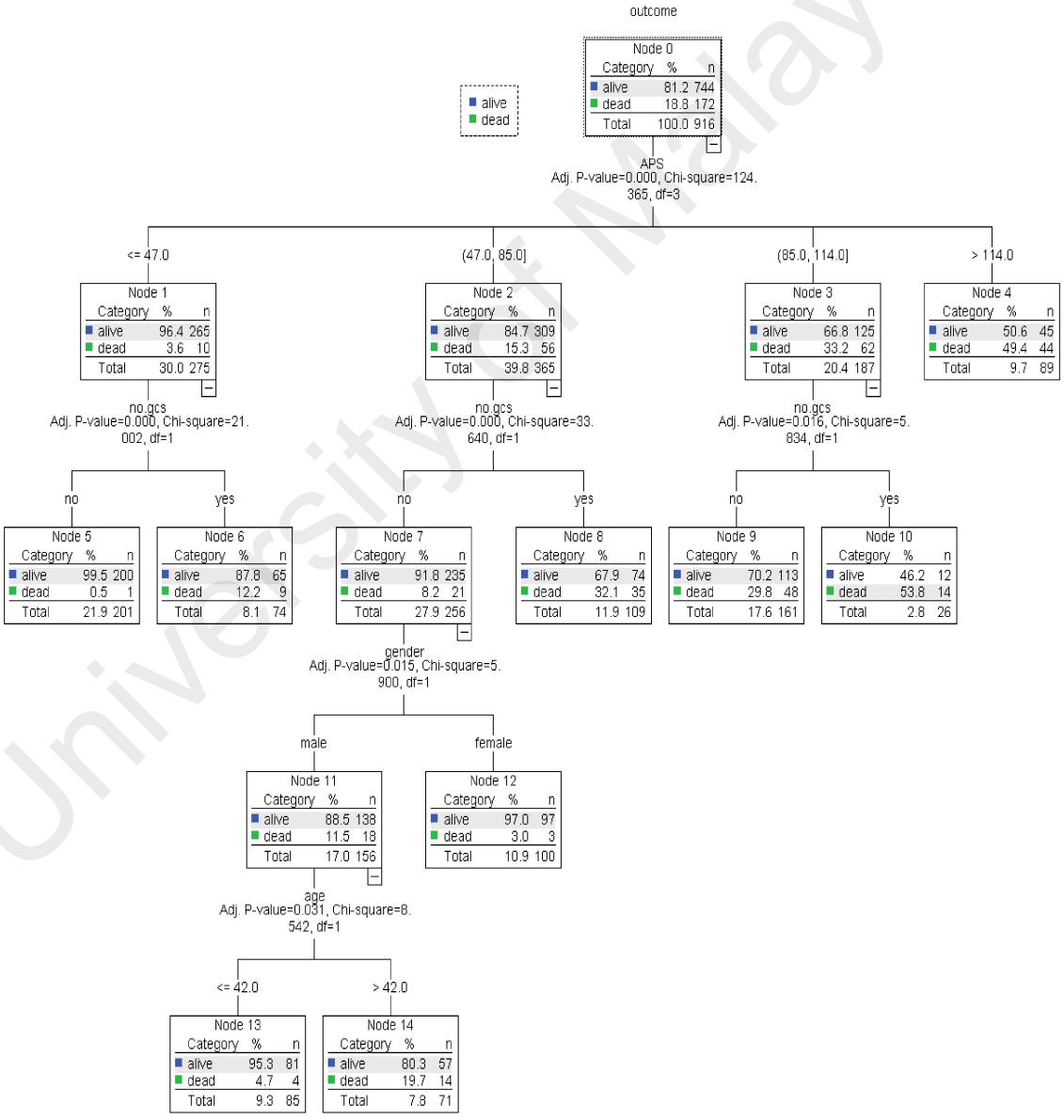


Figure 4.33: Decision tree based on *n*=916 patients (training cohort).

APS was the best discriminator between survivors and non-survivors in HSA ICU. The first level was divided into four decision rules according to APS values, i.e. APS ≤ 47.0, 47.0 < APS ≤ 85.0, 85.0 < APS ≤ 114.0 and APS >114.0. There was positive association between increasing APS and in-ICU mortality risk, where groups of patients with higher APS had higher in-ICU mortality risks. The associated mortality risks for each patient group were: APS ≤ 47.0 (3.6%), 47.0 < APS ≤ 85.0 (15.3%), 85.0 < APS ≤ 114.0 (33.2%) and APS >114.0 (49.4%).

At the second level in the decision tree, patients were stratified according to whether GCS score information was available or not. The first three APS groups (APS ≤ 47.0, 47.0 < APS ≤ 85.0 and 85.0 < APS ≤ 114.0) were sub-classified according to whether GCS score was available or not. Patients with high APS values and without GCS score were classified in the high-risk subgroups. Node 10 (subgroup 85.0 < APS ≤ 114.0 and without GCS score) recorded the highest probability of mortality at 53.8%. The intermediate risk groups included Node 8 (32.1%) and Node 9 (29.8%). On the other hand, the low risk groups were patients with low APS values (with or without GCS score information). The mortality risks for these subgroups were 0.5% (Node 5) and 12.2% (Node 6). Patients without GCS score had higher mortality risks compared to those with GCS score.

The third and fourth levels involved stratification of patients according to gender and age respectively for patients with 47.0 < APS ≤ 85.0 and with GCS score. Female patients in this subgroup (Node 12) had a much lower risk of mortality (3%). Male patients in this subgroup were further subdivided into two categories according to age. Older male patients (age > 42.0 years old, Node 14) had a higher probability of mortality (19.7%), whereas younger male patients (age ≤ 42.0, Node 13) were associated with a lower risk (4.7%).

The decision tree derived using the training cohort was then tested for its ability to risk stratify patients in the validation cohort that involved 195 patients. Figure 4.34 shows the corresponding decision tree that was generated using the validation cohort. The results obtained through the validation cohort were consistent to the results derived using the training cohort.

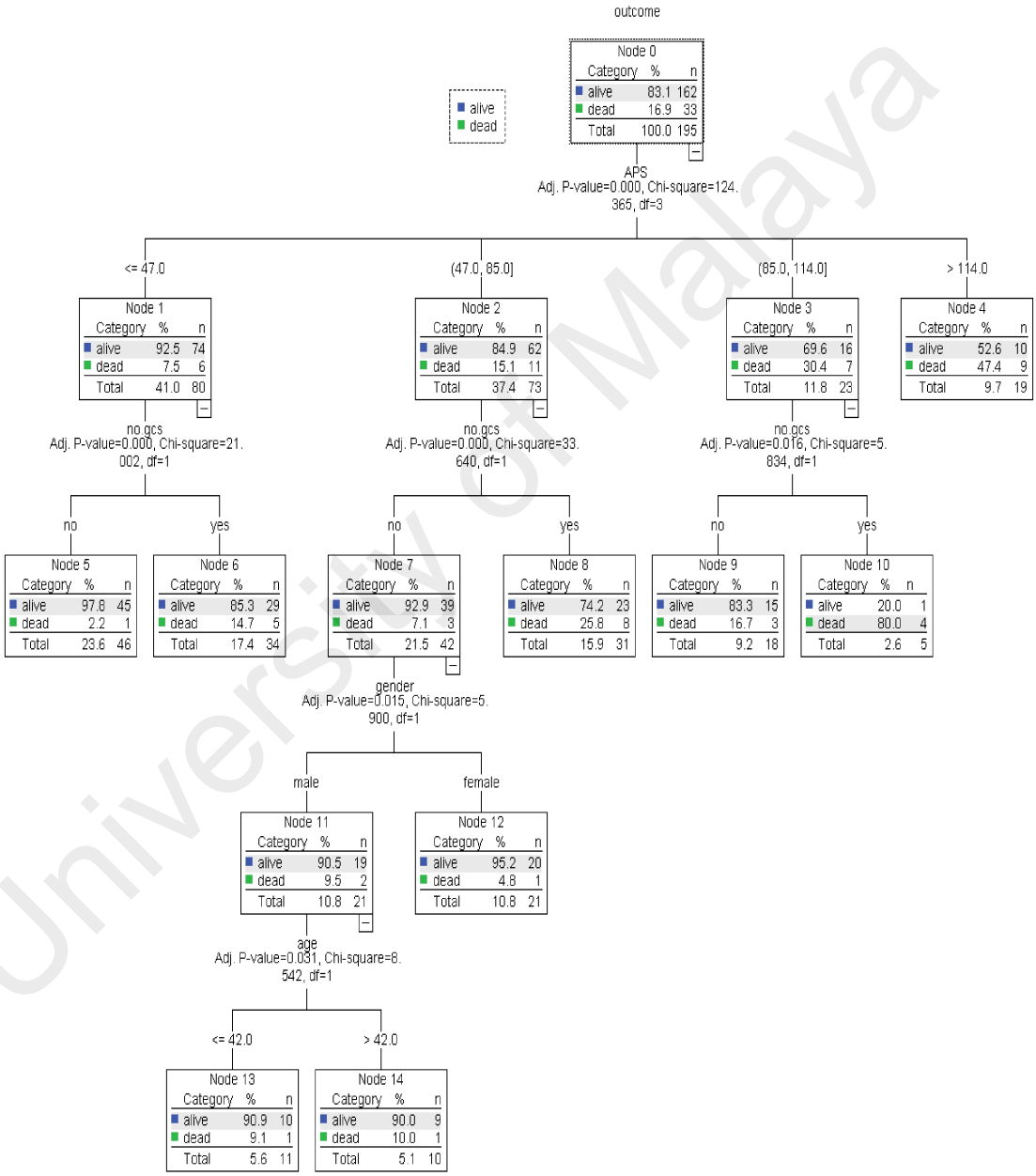
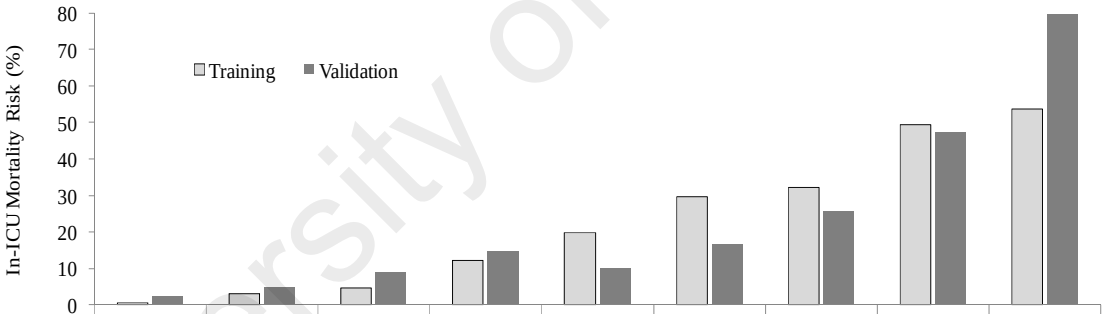


Figure 4.34: Decision tree based on *n*=195 patients (validation cohort).

The decision tree was able to stratify patients into three low-risk groups (Nodes 5, 12, 13), four intermediate-risk groups (Nodes 6, 14, 9, 8) and two high-risk groups (Nodes 4, 10). Figure 4.35 shows that the predicted in-ICU mortality risks in the validation cohort were higher than the training cohort in the three low-risk groups. In three of the four intermediate- risk groups (Nodes 14, 9, 8), predicted mortality rates were lower in the validation cohort compared to the training cohort. There were no substantial differences in the predicted mortality rates between validation and training cohorts in Node 4, which was considered as a high-risk group. On the other hand, a large discrepancy in the predicted mortality rates between the validation and training cohorts was observed in Node 10. However, the predictive accuracy of the validation cohort in this node was affected by the small number of patients in this subgroup.



	Node 5 Low	Node 12 Low	Node 13 Low	Node 6 Intermediate	Node 14 Intermediate	Node 9 Intermediate	Node 8 Intermediate	Node 4 High	Node 10 High
Training	0.5	3	4.7	12.2	19.7	29.8	32.1	49.4	53.8
Validation	2.2	4.8	9.1	14.7	10	16.7	25.8	47.4	80

Figure 4.35: Comparison of predicted in-ICU mortality risks in nine terminal nodes between training and validation cohorts.

Table 4.28 displays the classification accuracy of the decision trees that were derived using both the training and validation cohorts. The overall percentages of accurate predictions for the training and validation cohorts were rather high at 81.4% and 84.6% respectively. The overall estimated risk of death for the training cohort was

0.186 (standard error = 0.013), whereas the corresponding estimate for the validation cohort was 0.154 (standard error = 0.026).

Table 4.28: Classification accuracy of training and validation decision trees.

Cohort	Observed	Predicted		
		Alive	Dead	Percent Correct
Training	Alive	732	12	98.4%
	Dead	158	14	8.1%
	Overall Percentage	97.2%	2.8%	81.4%
Validation	Alive	161	1	99.4%
	Dead	29	4	12.1%
	Overall Percentage	97.4%	2.6%	84.6%

Table 4.29 shows the validation results of the decision tree using data set from 195 patients. The decision tree approach produced comparable results to the Bayesian logistic regression model in terms of discrimination and calibration. Discrimination was observed to be good, with AUC = 0.791 (see Figure 4.36 for ROC curve). The predicted and observed risks of mortality were compared in each of the nine nodes (see Figure 4.35). The decision tree approach produced overall good calibration (p-value > 0.05 in Hosmer and Lemeshow goodness-of-fit test). However, the Brier Score for the decision tree approach was not as satisfactory as the Brier Score obtained through the Bayesian approach. The higher Brier Score value indicated that the decision tree model was more non-informative compared to the Bayesian model. Overall, the decision tree approach produced a better SMR value of 0.946, which indicated close agreement between the overall observed and predicted mortality risks.

Table 4.29: Validation results of decision tree based on n=195 patients.

Performance indicator	Decision tree (validation)
AUC (95% confidence interval)	0.791 (0.728, 0.846)
Hosmer-Lemeshow statistic (p-value)	7.75 (p-value = 0.3352)
Brier Score	0.179
SMR (95% confidence interval)	0.946 (0.651, 1.328)

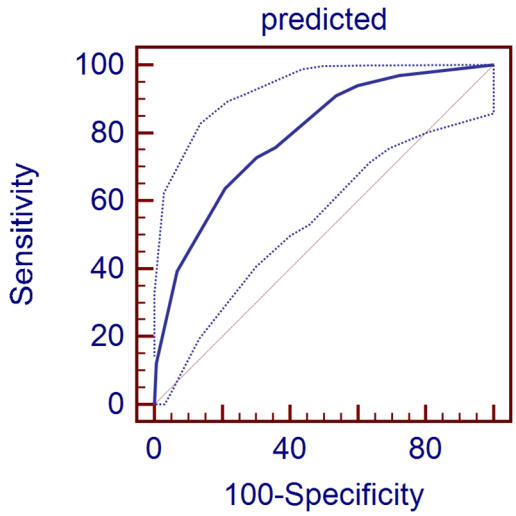


Figure 4.36: Receiver Operating Characteristic curve of Decision Tree (validation).

CHAPTER 5: DISCUSSION AND CONCLUSION

5.1 Discussion on performance of APACHE IV in HSA ICU

The first part of the study involved application of a well-established ICU prognostic model in a single-centre ICU at the Hospital Sultanah Aminah Johor Bahru (HSA ICU). The availability of numerous severity-of-illness scoring systems and mortality prediction models presented an initial problem in the choice of model to be emulated in this particular study. Over the years, these scoring and mortality systems have evolved through many generations, where the complexity of each generation was further increased through application of advanced statistical techniques and the latest technology in data acquisition.

The advantages and limitations of well-established ICU prognostic models were discussed in the literature review chapter. APACHE IV model was identified as the ideal reference/benchmark model in this study. However, this model was finally found to be not suitable for application in HSA ICU. Although APACHE IV exhibited acceptable discrimination, overall calibration and model fit was found to be quite poor. The APACHE IV model overestimated mortality risk in HSA ICU, especially for the mid- to high- risk ranges. The model's lack of fit was likely heavily influenced by differences in patient characteristics between APACHE IV and HSA ICU data sets, especially in terms of age, disease and admission types. Differences in ICU admission policies, management and practices between ICUs in the USA and Malaysia resulted in these differences in patient profiles. Admissions that were enrolled in the development of APACHE IV mostly involved elderly patients, where the mean age of patients was reportedly in the 60s range. On the other hand, admissions to HSA ICU were very much younger in age, where the mean age was observed to be in the 40s range. Approximately 30% of the total admissions in HSA ICU were younger than 30 years of

age. Further analysis revealed that a significant proportion of these younger patients were admitted due to trauma, where most of them were involved in motorcycle road accidents and were transferred from the Accident & Emergency (A&E) unit. This finding corroborated the national statistics of road injury and fatality involving motorcyclists in Malaysia (Abdul Manan & Várhelyi, 2012). This group of younger patients also contributed towards the higher percentage of emergency surgery admissions in HSA ICU.

Admissions to HSA ICU had higher severity-of-illness scores compared to the admissions used in development of APACHE IV. The mean APS in HSA ICU was considerably much higher than that of APACHE IV. This discrepancy could be caused by differences in the quality of treatment, infrastructure facilities and resources between ICUs in the USA and Malaysia. In addition, physiological components of patients were also partly affected by factors such as genetics composition, lifestyle, cultural and dietary habits. The higher APS values in the Malaysian cohort of patients could also be due to a higher variability in data in HSA ICU. This could inadvertently cause the possibility of extreme values being chosen as the worst values, and thus contributed to higher APS values. In this study, frequency of data collection followed the HSA ICU practice. Intervals of data collection were not equal time-spaced for all physiological variables. In particular, frequency of data collection was higher for routine variables that were easily available, compared to physiological variables that required laboratory assessments. The choice of worst values for the physiological variables was dependent on data availability. There was also the possibility that the choice of worst values for infrequently measured variables were affected by detection bias (Holmes et al., 2005).

Over the years, there were considerable debates regarding the use of the Standardised Mortality Ratio (SMR) as a valid and reasonable indicator of quality of care that was being provided in an ICU (Goldman & Brender, 2000; Jarman, 2008; van

Gestel et al., 2012). There were also concerns regarding the credibility of the SMR index as a screening mechanism to distinguish between low and high performance ICUs (Scott, Brand, Phelps, Barker, & Cameron, 2011). Although the SMR may not be a perfect indicator for evaluation of model performance and inter-ICUs comparison, it is still currently recognised as a universal standard reporting tool for hospital mortality, especially in countries such as the United Kingdom, United States, Australia, France, Canada, Japan, Hong Kong and Singapore (Jarman et al., 2010). Moreover, this index has been applied to evaluate model performance in many well-established hospital prognostic models, including APACHE IV. As such, reporting of SMR was still included in this study, although interpretation of the index should be done with caution.

External validation of APACHE IV in HSA ICU produced a low SMR value of 0.668. There were two possible interpretations for this finding. First, this low SMR value suggested poor calibration of APACHE IV, where the model overestimated risk of mortality in the Malaysian ICU. Alternatively, it could also indicate that the HSA ICU performed well with severely ill patients. In my opinion, the first interpretation seemed more reasonable than the second one. There were noticeable differences in case mix (age distribution and percentage of emergency surgery admissions) between the HSA ICU and APACHE IV cohorts. Furthermore, the proportions of patients in certain disease subgroups (eg. trauma, cardiovascular, respiratory, neurological) were higher than other subgroups (eg. genitourinary, musculoskeletal/skin) in the HSA ICU. These differences probably contributed towards the lower SMR value. In addition, the ICU was facing a shortage of trained personnel during the period of study. Coupled with the fact that Malaysia is still a developing country, where infrastructures and facilities may not be as well equipped as ICUs in the developed countries, it is highly improbable for the HSA ICU to be classified as having superior performance compared to the ICUs

enrolled in the APACHE IV study. Overall, this study has shown that the APACHE IV model was not suitable for application in HSA ICU, without further customisation.

5.2 Discussion on performance of Bayesian models

In recent years, advances in computing technology have led to an increase in the use of Bayesian techniques in statistical modelling and inference. Bayesian Markov Chain Monte Carlo (MCMC) and decision tree methods were identified as alternatives to the frequentist approach in the modelling of ICU mortality risk in this study. The Bayesian approach was chosen because it is suitable for high dimensional models and is able to accommodate small sample sizes. In addition, the Bayesian approach is appealing because it allows quantification of uncertainty in all parameters via probability and provides a more intuitive and meaningful inference of model parameters.

Using information and variables that were collected in APACHE IV, Bayesian MCMC approach was applied in the development of various combinations of multivariable logistic regression models that predict mortality risk in HSA ICU. Five main types (Type W, Type M, Type P, Type A and Type F) of Bayesian models were proposed and described in Chapter 3. Type W comprised the main effects model with APS variable. Variables that were included in Type W models were age, gender, Acute Physiology Score (APS), being on mechanical ventilation, absence of Glasgow Coma Scale (GCS) score due to sedation/paralysed, presence of chronic health, diabetes and ICU admission diagnoses.

The APS variable was a crucial predictor of ICU mortality risk in all four generations of the APACHE system. The APS component in APACHE system utilised scores that were assigned to the worst values of each physiological component because this approach produced better explanatory power (Knaus et al., 1991). However, this approach in assigning scores to the physiological variables also involved some elements

of subjectivity. Moreover, the approach in using worst values to represent the physiological components could be affected by detection bias and the frequency of data collection (Holmes et al., 2005). An important contribution of this research is in finding alternative treatments to the APS variable in the predictive models. In order to investigate the effect and importance of APS variable, alternative approaches in the modelling of physiological components were explored as substitutes for the APS variable in the other four types of models (Types M, P, A and F).

In the second model type, the effect of using mean values of physiological observations instead of the worst values was investigated. Type M model comprised the main effects model with a new variable known as APS.mean. The APS.mean variable was calculated by combining scores that were assigned to the mean values of the physiological variables, instead of the worst values. The results in this study indicated that the ability of APS.mean variable in explaining risk of death in HSA ICU was much inferior compared to the APS variable. Model performance was not improved with the use of APS.mean variable and that it was still better to retain the original APS calculation approach.

On the other hand, the methods applied in Type P, Type A and Type F models allowed the use of all available physiological observations and eliminated the need for a scoring component such as APS. In Type P models, the APS variable was removed and substituted with dichotomous classifications (normal/abnormal) for each of the worst values of the physiological variables. Classification of abnormality for the physiological variables was based on the APACHE IV definitions. The DIC values in Type P models were much higher than Type W models. The high DIC values could probably be caused by the increased complexity of the Type P models, where the total variables in the Type P models were twice the number of variables in Type W models. Nevertheless, a difference of more than 20 units in DIC values between the two model types greatly

favoured the use of APS variable in Type W models as better options compared to the dichotomous classifications for worst physiological variables in Type P models.

The effect in using frequency of abnormal physiological observations throughout the first day of ICU admission was explored in the Type A models. Two methods were proposed for the calculation of the frequency of abnormal physiological observations. The first method involved the introduction of a new single variable to replace APS in quantifying the physiological components models A1 and A2. This variable was calculated as the simple ratio of the number of abnormal physiological variables over the total number of physiological variables within the first day of ICU admission for each patient. Although both models A1 and A2 exhibited good discrimination and calibration, their DIC values were much higher compared to models W1 and W2 respectively. This implied that there was no improvement in model fit when the APS variable was replaced with the new variable introduced in models A1 and A2. Conversely, both models A1 and A2 outperformed the Type M models with lower DIC values. These findings suggested that the use of APS.mean in Type M models was less effective than the frequency of abnormal physiological observations in models A1 and A2.

The second method in Type A models required direct incorporation of the frequency of abnormal observations for each individual physiological variable in models A3 and A4. The approach used in models A3 and A4 was rather similar to models P1 and P2. Instead of using dichotomous classifications (normal/abnormal) for each of the worst values of the physiological variables in models P1 and P2, frequency of abnormal observations for each individual physiological variable was used in models A3 and A4. Although models A3 and A4 slightly edged models P1 and P2 in terms of discrimination, overall model fit and calibration across different subgroups of patients

were much lacking in models A3 and A4. These results indicated that the performances of models P1 and P2 were generally better compared to models A3 and A4.

In the fifth type of models (Type F), factor analysis approach was applied to introduce a latent variable that represented the physiological components. This involved calculation of factor scores for the standardised worst values of the physiological variables within the first day of ICU admission. The performances of the Type F models were the worst among the five different types of models. The DIC values in Type F models were the highest among the five different types of models. Moreover, Type F models also had the worst discrimination and calibration properties compared to the other model types. Application of latent variables using factor analysis approach was not able to explain the physiological components adequately, compared to the other methods used in the other types of models.

In summary, a comparison of the DIC values among the five types of models indicated that overall model fit was the best in Type W models. Based on the comparison of DIC values, the types of models were rated from best to worse in the order of Type W > Type P > Type A > Type M > Type F. The Type W models with APS outperformed all the other model types, where APS was proven to be a good measure of severity of illness. The ability of the APS in explaining ICU mortality risk in HSA ICU indicated that the APS variable remained a relevant and significant risk factor in predicting mortality risk in the Malaysian cohort of patients.

Among the Type W models, Model W1 was identified and recommended as the best model to be implemented in HSA ICU. Detailed descriptions and results of this model were presented and discussed in Chapter 4. The model contained seven variables (age, gender, APS, absence of GCS score due to patient being sedated/paralysed, mechanical ventilation, presence of chronic health and ICU admission diagnoses) that are readily available in any intensive care unit setting. Gender, APS and absence of

GCS score were found to be significant determinants of HSA ICU mortality risk. The odds of dying for female patients were 53% less compared to male patients. The percentage of patients without GCS score in HSA ICU was rather high at slightly more than 20%. The odds of dying for these patients without GCS score was approximately five to six times higher than patients with GCS score. Increasing APS was also associated with a higher ICU mortality risk, where the odds of dying increased by 4% for every increase of one unit in APS. Patients on mechanical ventilation or with chronic health were also associated with higher risks of dying. Although age, mechanical ventilation, presence of chronic health and ICU admission diagnoses were not statistically significant based on the 95% credible intervals of the posterior means in model W1, they were still included due to their clinical importance.

This study found no significant correlation between age and the probability of mortality. There was also no positive association between age and APS. Patients who died were not necessarily older patients with higher APS as the cohort of patients in HSA ICU included a large proportion of young patients with high APS values. This conclusion was also consistent with the outcomes observed in another study in a Singapore ICU, where it was reported that age was not a significant factor in determining mortality outcome (Lee, Hui, Lim, & Tan, 1993). Although age was found to have no effect modification when assessed across other variables in model W1, this variable was still included due to its clinical relevance.

Despite being a multiracial country, ethnicity was not a significant predictor of death in this study. One possible explanation for this could be that Malaysians generally have similar dietary and eating habits although they come from culturally diverse backgrounds. Moreover, the increasing number of inter-ethnic marriages over the years could have also contributed towards better integration and assimilation of cultural values and lifestyles among the various ethnic groups in Malaysia. Although the

percentage of diabetic patients in the HSA ICU was considered quite high at approximately 20%, diabetes was not statistically significant when combined with other variables in model W2. This finding resonated with the theory that although diabetic patients were susceptible to more complications, diabetes was not associated with increased in-ICU mortality risk (Siegelaar, Devries, & Hoekstra, 2010; Siegelaar, Hickmann, Hoekstra, Holleman, & DeVries, 2011; Vincent, Preiser, Sprung, Moreno, & Sakr, 2010). This study also revealed that there were no substantial differences in the prevalence of diabetes among the three major ethnic groups (Malay, Chinese and Indian) of patients in HSA ICU.

The results in Section 4.4.2 indicated that the Bayesian and frequentist (MLE) methods produced results that were quite close in agreement and provided similar conclusions in terms of performance in model W1. This was probably due to the data set being sufficiently large enough, especially for the MLE approach. Furthermore, a large number of iterations was employed in the Bayesian MCMC simulations in order to ensure model convergence. Although absence of prior experience necessitated the use of non-informative (vague) prior distributions in the models, the Bayesian approach was able to produce results that were somewhat consistent with the frequentist method. This was one of the advantages of the Bayesian approach in allowing the data to speak for themselves. The ability to quantify uncertainty in the model parameters also provided more flexibility in the Bayesian modelling approach. The results in this study indicated that the Bayesian approach produced smaller standard errors and narrower credible intervals compared to the frequentist (MLE) approach. In addition, comparison of the deviance values also revealed that better model fit was achieved through the Bayesian approach.

5.3 Discussion on performance of decision tree model

The decision tree approach was also presented as an alternative method in the modelling of ICU mortality risk in this study. The construction of the decision tree was based on variables defined in model W1, with risk of death upon discharge from ICU being the outcome variable. Variables that were significant in the Bayesian model (gender, APS, absence of GCS score) were also similarly found to be significant predictors in the decision tree model. However, age was also observed to be significant and was included in the decision tree model. Presence of chronic health, mechanical ventilation and ICU admission diagnoses were not significant and were excluded from the decision tree model.

Recursive partitioning of the training cohort identified the APS as the best single predictor of mortality risk in HSA ICU, where positive association was detected between increasing APS and ICU mortality risk. This finding suggested that APS had the strongest explanatory power in predicting mortality risk in HSA ICU and supported the importance of APS variable in stratifying patients according to their mortality risks. Absence of GCS score was the second best predictor in the decision tree model, where patients without GCS score were associated with higher mortality risks than patients with GCS score. Gender was placed in the third tier of the decision tree model. Consistent with results obtained in the Bayesian model, the decision tree model also indicated that female patients had lower risk of mortality compared to male patients. In the fourth tier, male patients with moderate APS values between 48 and 85 and without GCS score were then divided into two subgroups according to their age. The decision tree model stratified patients into a total of nine groups (three low-risk, four intermediate-risk and two high-risk), with mortality ranging from 0.5% to 53.8%.

The area under receiver operating characteristic curve in the decision tree model (AUC=0.791, 95% CI: (0.728, 0.846)) was comparable to the Bayesian model

(AUC=0.810, 95% CI: (0.748, 0.862)). Predicted mortality was slightly higher than actual mortality in the low risk-groups, whereas the decision tree model slightly over predicted mortality in the intermediate risk-groups. A large discrepancy between predicted and actual mortality in the last high-risk group was due to the small number of patients in this subgroup. Although the decision tree model provided a simple and easy-to-use mortality prediction tool, its higher Brier Score indicated that the decision tree model was not as informative as the more complicated Bayesian logistic regression model. Application of the decision tree model is also restricted to prediction of group mortality and is not suitable for prediction of individual mortality risk. In addition, the decision tree model is particularly sensitive to small changes in the sample size of the training cohort. Prediction accuracy in a decision tree model is also affected by a small training sample because the number of available training examples in each branch decreases exponentially as the decision tree is being built (Katz, Shabtai, Rokach, & Ofek, 2014). These limitations favoured the option of using the Bayesian model over the decision rule model in predicting mortality risk in this study.

5.4 Research Limitations

There were several limitations to this study. First, due to the single-centre nature of this study, findings that were obtained in HSA ICU may not be representative of other ICUs in Malaysia. This study has shown that differences in case mix, clinical practice, admission and discharge policies, quality of treatment and care being provided were important factors that affected the suitability and applicability of a model in a different setting. It is important to note that although the proposed model in this study was not meant to be generalised to other ICUs, differences in case mix, clinical settings and management policies are potential factors that limit generalisation of the models proposed in this study to other ICUs.

The focus of this study was in predicting in-ICU mortality risk, and not in-hospital deaths. Critically ill patients that are discharged from the ICU to a palliative care facility or high dependency unit may appear as a successful discharge. However, in reality, these discharge cases may be very similar to an in-hospital death. In addition, the proposed model was also not able to take into account cases of patients dying immediately upon discharge from ICU. In this sense, a model to predict death within a year of discharge from the ICU could be a useful extension that can be done in future research. Although these events are relatively infrequent, they have a potential to impact on the results, given the relatively small number of deaths in the ICU. In addition, patients with multiple admissions and very long stays were excluded according to the exclusion criteria defined in APACHE IV. Although exclusion of these patients may make modelling easier, they reduce the generalisability of the model to wider ICU patient population.

Prediction of mortality risk in this study was based on data that were collected on the first day of ICU admission. One of the problems that were encountered in the data collection process was missing data. Patients with missing or incomplete data were excluded from the study, giving rise to an overall smaller data set. The sample size in this single centre study could be considered as relatively small compared to other large-scale studies that involve multiple ICUs and institutions. This issue of small sample size presented a problem in that the case mix was probably not sufficiently diverse. This was also a limiting factor in the analysis of uniformity of fit among different subgroups in this study.

There was also the issue that the sample size may have influenced the Hosmer-Lemeshow goodness-of-fit test statistic, since the Hosmer-Lemeshow test is known to be sensitive to sample size when the model deviates from perfect fit. It was a concern that a small sample size may not detect poor fit, even when there was actual deviation

from perfect fit. Hosmer and Lemeshow (2000) recommended that sample sizes should ideally be more than 400 for application of the Hosmer-Lemeshow goodness-of-fit test. In their work to evaluate effective sample sizes for external validation studies that involve prediction logistic regression models, Vergouwe et al. (2005) suggested a minimum of 100 events and 100 non-events for external validation samples. The sample size used in this study fulfilled both recommendations and should be adequate in providing reasonable power to detect overall model fit. Moreover, the Hosmer-Lemeshow test was just only one of the criteria used in assessing model calibration in this study. This method was also supplemented with calibration curves that assessed difference between observed and predicted mortality outcomes across different risk ranges.

Although a multi-centre study will probably mitigate issues regarding sample size and case mix, it was unfortunate this aspiration was hindered by lack of cooperation and commitment from other ICUs in the country. It is hoped that with the publication of the results obtained in this study, other ICUs in Malaysia will be more agreeable to participate in a larger scale multi-centre study in the future. In addition, future research can also be done to compare the performances of other models such SAPS 3 Admission Score model or MPM_0-III admission model in the Malaysian cohort of patients. However, implementation of these models is still subject to availability of resources and funds.

5.5 Concluding remarks

In conclusion, this study has shown that APACHE IV model was not suitable for application in HSA ICU because the model overestimated mortality risk in the Malaysian ICU. In order to address this problem, the Bayesian Markov Chain Monte Carlo and decision tree approaches were chosen as alternative modelling strategies in

developing a suitable model that can be used in HSA ICU. Despite being more complicated, the Bayesian MCMC approach was able to offer a different perspective in the analysis and modelling of in-ICU mortality risk. Although this approach has been mainly used to generate mortality predictions for single specific diseases, this study has shown that its application can be successfully expanded to include mortality prediction for a cohort of critically ill patients with multiple presenting symptoms. In this sense, the methodological aspects of this research can also be viewed as contribution to the existing literature. The decision tree approach was also able to produce predictions that were comparable to the Bayesian models. However, a main disadvantage of the decision tree model is that it is not as informative as the Bayesian models and is not suitable for predicting individual mortality risk. Moreover, the decision tree model requires a large training set in order to achieve good prediction accuracy. Hence, application of Bayesian model is recommended over the decision rule model in this study.

There is immense potential in using prognostic models to enhance the quality of critical care in Malaysia. To the best of the author's knowledge, at the time of writing, this was the first study that involved external validation of APACHE IV, as well as, application of Bayesian approach in modelling ICU mortality risk in Malaysia. The study was instrumental in providing an insight on the characteristics of patients who were admitted to a Malaysian ICU. It is hoped that the recommended model in this study may serve as a clinical decision support tool that complements physician assessment in the long run. Moreover, the proposed model can also be utilised to monitor the performance of the ICU over time and offer guidance in terms of resource planning and allocation of beds in the ICU. Finally, periodic recalibration of the proposed model should be conducted from time to time as it is a well-known fact that all prognostic models have the tendency to deteriorate in performance over time.

REFERENCES

Abdul Manan, M. M., & Várhelyi, A. (2012). Motorcycle fatalities in Malaysia. *IATSS Research, 36*(1), 30-39. doi: http://dx.doi.org/10.1016/j.iatssr.2012.02.005.

Abu-Hanna, A., & de Keizer, N. (2003). Integrating classification trees with local logistic regression in Intensive Care prognosis. *Artificial Intelligence in Medicine, 29*(1-2), 5-23. doi: S0933365703000472 [pii].

Aegerter, P., Boumendil, A., Retbi, A., Minvielle, E., Dervaux, B., & Guidet, B. (2005). SAPS II revisited. *Intensive Care Medicine, 31*(3), 416-423. doi: 10.1007/s00134-005-2557-9.

Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ, 338*, b605.

Anderson, R. P., Jin, R., & Grunkemeier, G. L. (2003). Understanding logistic regression analysis in clinical reports: an introduction. *The Annals of Thoracic Surgery, 75*(3), 753-757. doi: S0003-4975(02)04683-0 [pii].

Arabi, Y., Haddad, S., Goraj, R., Al-Shimemeri, A., & Al-Malik, S. (2002). Assessment of performance of four mortality prediction systems in a Saudi Arabian intensive care unit. *Critical Care, 6*(2), 166-174.

Badriyah, T., Briggs, J. S., & Prytherch, D. R. (2012). Decision Trees for Predicting Risk of Mortality using Routinely Collected Data. *World Academy of Science, Engineering and Technology, 62*, 660-663.

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology, 54*(10), 979-985. doi: S0895-4356(01)00372-9 [pii].

Balaji, B., Rao, A. B., Kumar, V. S., & Sammaiah. (2016). Performance of simplified acute physiology score 3 admission score as a predictor of ICU mortality in a tertiary care hospital of rural Telangana, India. *International Journal of Advances in Medicine, 3*(3), 716-720. doi:10.18203/2349-3933.ijam20162523.

Bastos, P. G., Sun, X., Wagner, D. P., Knaus, W. A., & Zimmerman, J. E. (1996). Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. *Intensive Care Medicine, 22*(6), 564-570.

Becker, R. B., Zimmerman, J. E., Knaus, W. A., Wagner, D. P., Seneff, M. G., Draper, E. A. (1995). The use of APACHE III to evaluate ICU length of stay, resource use, and mortality after coronary artery by-pass surgery. *Journal of Cardiovascular Surgery (Torino), 36*(1), 1-11.

Bedrick, E. J., Christensen, R., & Johnson, W. (1997). Bayesian Binomial Regression: Predicting Survival at a Trauma Center. *The American Statistician, 51*, 211-218.

Bertolini, G., D'Amico, R., Apolone, G., Cattaneo, A., Ravizza, A., Iapichino, G. (1998). Predicting outcome in the intensive care unit using scoring systems: is new better? A comparison of SAPS and SAPS II in a cohort of 1,393 patients. GiViTi Investigators (Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva). Simplified Acute Physiology Score. *Medical Care, 36*(9), 1371-1382.

Bhattacharyya, M., & Todi, S. (2009). APACHE IV: benchmarking in an Indian ICU. *Critical Care, 13*(Suppl 1), P510.

Blanco, R., Inza, I., Merino, M., Quiroga, J., & Larranaga, P. (2005). Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics, 38*(5), 376-388. doi: 10.1016/j.jbi.2005.05.004.

Boyd, O., & Grounds, M. (1994). Can standardized mortality ratio be used to compare quality of intensive care unit performance? *Critical Care Medicine, 22*(10), 1706-1709.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont California: Wadsworth, Inc.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1-3.

Brinkman, S., Bakhshi-Raiez, F., Abu-Hanna, A., de Jonge, E., Bosman, R. J., Peelen, L. (2011). External validation of Acute Physiology and Chronic Health Evaluation IV in Dutch intensive care units and comparison with Acute Physiology and Chronic Health Evaluation II and Simplified Acute Physiology Score II. *Journal of Critical Care, 26*(1), 105 e111-108. doi: S0883-9441(10)00184-X [pii]10.1016/j.jcrc.2010.07.007.

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434-455. doi: 10.1080/10618600.1998.10474787.

Capuzzo, M., Moreno, R. P., & Le Gall, J.-R. (2008). Outcome prediction in critical care: the Simplified Acute Physiology Score models. *Current Opinion in Critical Care, 14*(5), 485-490.

Castella, X., Artigas, A., Bion, J., & Kari, A. (1995). A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European/North American Severity Study Group. *Critical Care Medicine, 23*(8), 1327-1335.

Chen, M.-H., Ibrahim, J. G., & Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*, 223-242.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association, 74*(368), 829-836. doi: 10.1080/01621459.1979.10481038.

Cole, S. R., Chu, H., & Greenland, S. (2014). Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology, 179*(2), 252-260. doi: 10.1093/aje/kwt245.

Congdon, P. (2001). *Bayesian Statistical Modelling.* Chichester: John Wiley & Sons.

Costello, A. B., & Osborne, J. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *10*, 1-9.

Cook, D. A. (2006). Methods to assess performance of models estimating risk of death in intensive care patients: a review. *Anaesthesia and Intensive Care, 34*(2), 164-175. doi: 2005066 [pii].

Cook, D. A., Joyce, C. J., Barnett, R. J., Birgan, S. P., Playford, H., Cockings, J. G. (2002). Prospective independent validation of APACHE III models in an Australian tertiary adult intensive care unit. *Anaesthesia and Intensive Care, 30*(3), 308-315.

Costa e Silva, V. T., de Castro, I., Liano, F., Muriel, A., Rodriguez-Palomares, J. R., & Yu, L. (2011). Performance of the third-generation models of severity scoring systems (APACHE IV, SAPS 3 and MPM-III) in acute kidney injury critically ill patients. *Nephrology Dialysis Transplantion, 26*(12), 3894-3901. doi: gfr201 [pii]10.1093/ndt/gfr201.

Cowen, J. S., & Kelley, M. A. (1994). Errors and bias in using predictive scoring systems. *Critical Care Clinics, 10*(1), 53-72.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman and Hall.

Czepiel, S. A. (2002). Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. Retrieved from http://czep.net.

de Rooij, S. E., Abu-Hanna, A., Levi, M., & de Jonge, E. (2007). Identification of high-risk subgroups in very elderly intensive care unit patients. *Critical Care, 11*(2), R33.

Dragsted, L., Jorgensen, J., Jensen, N. H., Bonsing, E., Jacobsen, E., Knaus, W. A. (1989). Interhospital comparisons of patient outcome from intensive care: importance of lead-time bias. *Critical Care Medicine, 17*(5), 418-422.

Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology, 153*(12), 1222-1226.

Ebell, M. H. (2007). Predicting mortality risk in patients with acute exacerbations of heart failure. *American Family Physician, 75*(8), 1231-1233.

Escarce, J. J., & Kelley, M. A. (1990). Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. *Journal of the American Medical Association, 264*(18), 2389-2394.

Fischler, L., Lelais, F., Young, J., Buchmann, B., Pargger, H., & Kaufmann, M. (2007). Assessment of three different mortality prediction models in four well-defined critical care patient groups at two points in time: a prospective cohort study. *European Journal of Anaesthesiology, 24*(8), 676-683. doi: S026502150700021X [pii]10.1017/S026502150700021X.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association, 85*, 398-409.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo In Practice* (pp. 131-143). London: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*(4), 457-472. doi: 10.1214/ss/1177011136.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721 - 741.

Gerds, T. A., Cai, T., & Schumacher, M. (2008). The Performance of Risk Prediction Models. *Biometrical Journal, 50*(4), 457-479. doi: 10.1002/bimj.200810443.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid & J. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 169-193): Oxford University Press.

Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science, 7*(4), 473-483. doi: 10.1214/ss/1177011145.

Glance, L. G., Osler, T. M., & Dick, A. (2002). Rating the quality of intensive care units: is it a function of the intensive care unit scoring system? *Critical Care Medicine, 30*(9), 1976-1982.

Goldhill, D. R., & Withington, P. S. (1996). The effect of casemix adjustment on mortality as predicted by APACHE II. *Intensive Care Medicine, 22*(5), 415-419.

Goldman, D. A., & Brender, J. D. (2000). Are standardized mortality ratios valid for public health data analysis? *Statistics in Medicine, 19*(8), 1081-1088. doi: 10.1002/(SICI)1097-0258(20000430)19:8<1081::AID-SIM406>3.0.CO;2-A [pii].

Goodenough, D. J., Rossmann, K., & Lusted, L. B. (1974). Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology, 110*(1), 89-95. doi: 10.1148/110.1.89.

Greenland, S. (1995). Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology, 6*(4), 450-454.

Grunkemeier, G. L., & Jin, R. (2001). Receiver operating characteristic curve analysis of clinical risk models. *Annals of Thoracic Surgery, 72*(2), 323-326. doi: S0003-4975(01)02870-3 [pii].

Gunning, K., & Rowan, K. (1999). Outcome data and scoring systems. *British Medical Journal, 319*(7204), 241-244. doi: 10.1136/bmj.319.7204.241.

Hamra, G., MacLehose, R., & Richardson, D. (2013). Markov Chain Monte Carlo: an introduction for epidemiologists. *International Journal of Epidemiology, 42*(2), 627-634. doi: 10.1093/ije/dyt043.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97-109.

Heidelberger, P., & Welch, P. D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research, 31*(6), 1109-1144. doi: 10.1287/opre.31.6.1109.

Hernandez, A. M. R., & Palo, J. E. M. (2014). Performance of the SAPS 3 admission score as a predictor of ICU mortality in a Philippine private tertiary medical center intensive care unit. *Journal of Intensive Care, 2*(1), 29. doi: 10.1186/2052-0492-2-29.

Hess, K. R., Abbruzzese, M. C., Lenzi, R., Raber, M. N., & Abbruzzese, J. L. (1999). Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. *Clinical Cancer Research, 5*(11), 3403-3410.

Higgins, T. L., Kramer, A. A., Nathanson, B. H., Copes, W., Stark, M., & Teres, D. (2009). Prospective validation of the intensive care unit admission Mortality Probability Model (MPM0-III). *Critical Care Medicine, 37*(5), 1619-1623. doi: 10.1097/CCM.0b013e31819ded31.

Higgins, T. L., Teres, D., Copes, W. S., Nathanson, B. H., Stark, M., & Kramer, A. A. (2007). Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Critical Care Medicine, 35*(3), 827-835.

Higgins, T. L., Teres, D., & Nathanson, B. (2008). Outcome prediction in critical care: the Mortality Probability Models. *Current Opinion in Critical Care, 14*(5), 498-505.

Holmes, C. L., Gregoire, G., & Russell, J. A. (2005). Assessment of Severity of Illness. In J. B. Hall, G. A. Schmidt & L. D. H. Wood (Eds.), *Principles of Critical Care* (3rd ed., pp. 63 - 78). Blacklick, OH, USA: McGraw-Hill Professional Publishing.

Hosmer, D. W., & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics, A10*, 1043-1069.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second Edition ed.). New York: Wiley.

Hu, Y., Zhang, X., Liu, Y., Yan, J., Li, T., & Hu, A. (2013). APACHE IV Is Superior to MELD Scoring System in Predicting Prognosis in Patients after Orthotopic Liver Transplantation. *Clinical and Developmental Immunology, 2013*, 5. doi: 10.1155/2013/809847.

Hunt, D. L., Haynes, R. B., Hanna, S. E., & Smith, K. (1998). Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *Journal of the American Medical Association, 280*(15), 1339-1346.

Jae, W. C., Young, S. P., Young, S. L., Yeon, H. P., Chaeuk, C., Dong, P.,..., Jae, M. (2017). The Ability of the Acute Physiology and Chronic Health Evaluation (APACHE) IV Score to Predict Mortality in a Single Tertiary Hospital. *32*(3), 275-283. doi: 10.4266/kjccm.2016.00990.

Jarman, B. (2008). In defence of the hospital standardized mortality ratio. *HealthcarePapers, 8*(4), 37-42; discussion 69-75.

Jarman, B., Pieter, D., van der Veen, A. A., Kool, R. B., Aylin, P., Bottle, A. (2010). The hospital standardised mortality ratio: a powerful tool for Dutch hospitals to assess their quality of care? *Quality and Safety in Health Care, 19*(1), 9-13. doi: 10.1136/qshc.2009.032953.

Jeong, I., Kim, M., & Kim, J. (2003). Predictive accuracy of severity scoring system: a prospective cohort study using APACHE III in a Korean intensive care unit. *International Journal of Nursing Studies, 40*, 219-226.

Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine, 130*(6), 515-524.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 29*(2), 119-127.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795.

Katsaragakis, S., Papadimitropoulos, K., Antonakis, P., Strergiopoulos, S., Konstadoulakis, M. M., & Androulakis, G. (2000). Comparison of Acute Physiology and Chronic Health Evaluation II (APACHE II) and Simplified Acute Physiology Score II (SAPS II) scoring systems in a single Greek intensive care unit. *Critical Care Medicine, 28*(2), 426-432.

Katz, G., Shabtai, A., Rokach, L., & Ofek, N. (2014). ConfDTree: A Statistical Method for Improving Decision Trees. *Journal of Computer Science and Technology, 29*(3), 392-407. doi: 10.1007/s11390-014-1438-5.

Kay, R., & Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika, 74*(3), 495-501. doi: 10.1093/biomet/74.3.495.

Kazembe, L. N., Chirwa, T. F., Simbeye, J. S., & Namangale, J. J. (2008). Applications of Bayesian approach in modelling risk of malaria-related hospital mortality. *BMC Medical Research Methodology, 8*, 6. doi: 1471-2288-8-6 [pii] 10.1186/1471-2288-8-6.

Kherallah, M., Hazza, M., Dahhan, T., Tantawy, T., Jamil, M. G., & Al-Tarifi, A. (2008). Performance of the Acute Physiology and Chronic Health Evaluation IV at a Tertiary Saudi Hospital. *Chest, 134*(4), p112003.

Kim, H., & Loh, W.-Y. (2001). Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association, 96*(454), 589-604. doi: 10.1198/016214501753168271.

Kim, S., Kim, W., & Park, R. W. (2011). A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthcare Informatics Research, 17*(4), 232-243. doi: 10.4258/hir.2011.17.4.232.

King, E. N., & Ryan, T. P. (2002). A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression. *The American Statistician, 56*(3), 163-170. doi: 10.1198/00031300283.

Knaus, W. A. (2002). APACHE 1978-2001: the development of a quality assurance system based on prognosis: milestones and personal reflections. *Archives of Surgery, 137*(1), 37-41. doi: ssa1013 [pii].

Knaus, W. A., Draper, E. A., Wagner, D. P., & Zimmerman, J. E. (1985). APACHE II: a severity of disease classification system. *Critical Care Medicine, 13*(10), 818-829.

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G. (1991). The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest, 100*(6), 1619-1636.

Knaus, W. A., Zimmerman, J. E., Wagner, D. P., Draper, E. A., & Lawrence, D. E. (1981). APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical Care Medicine, 9*(8), 591-597.

Kramer, A. A. (2005). Predictive mortality models are not like fine wine. *Critical Care, 9*(6), 636-637.

Kuzniewicz, M. W., Vasilevskis, E. E., Lane, R., Dean, M. L., Trivedi, N. G., Rennie, D. J. (2008). Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *Chest, 133*(6), 1319-1327. doi: chest.07-3061 [pii]10.1378/chest.07-3061.

LaValley, M. P. (2008). Logistic Regression. *Circulation, 117*(18), 2395-2399. doi: 10.1161/circulationaha.106.682658.

Le Gall, J. R., Lemeshow, S., & Saulnier, F. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association, 270*(24), 2957-2963.

Le Gall, J. R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Mathieu, D. (1984). A simplified acute physiology score for ICU patients. *Critical Care Medicine, 12*(11), 975-977.

Le Gall, J. R., Neumann, A., Hemery, F., Bleriot, J. P., Fulgencio, J. P., Garrigues, B. (2005). Mortality prediction using SAPS II: an update for French intensive care units. *Critical Care, 9*(6), R645-652. doi: cc3821 [pii] 10.1186/cc3821.

Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution, 2*(2), 155-162. doi: 10.1111/j.2041-210X.2010.00063.x.

Ledoux, D., Canivet, J. L., Preiser, J. C., Lefrancq, J., & Damas, P. (2008). SAPS 3 admission score: an external validation in a general intensive care population. *Intensive Care Medicine, 34*(10), 1873-1877. doi: 10.1007/s00134-008-1187-4.

Lee, K. H., Hui, K. P., Lim, T. K., & Tan, W. C. (1993). Acute physiology and chronic health evaluation (APACHE II) scoring in the Medical Intensive Care Unit, National University Hospital, Singapore. *Singapore Med J, 34*(1), 41-44.

Lemeshow, S., Klar, J., Teres, D., Avrunin, J. S., Gehlbach, S. H., Rapoport, J. (1994). Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective, multicenter study. *Critical Care Medicine, 22*(9), 1351-1358.

Lemeshow, S., Teres, D., Avrunin, J. S., & Gage, R. W. (1988). Refining intensive care unit outcome prediction by using changing probabilities of mortality. *Critical Care Medicine, 16*(5), 470-477.

Lemeshow, S., Teres, D., Avrunin, J. S., & Pastides, H. (1987). A comparison of methods to predict mortality of intensive care unit patients. *Critical Care Medicine, 15*(8), 715-722.

Lemeshow, S., Teres, D., Klar, J., Avrunin, J. S., Gehlbach, S. H., & Rapoport, J. (1993). Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *Journal of the American Medical Association, 270*(20), 2478-2486.

Lemeshow, S., Teres, D., Pastides, H., Avrunin, J. S., & Steingrub, J. S. (1985). A method for predicting survival and mortality of ICU patients using objectively derived weights. *Critical Care Medicine, 13*(7), 519-525.

Letchumanan, G. R., Wan Nazaimoon, W. M., Wan Mohamad, W. B., Chandran, L. R., Tee, G. H., Jamaiyah, Y.,..., Ahmad Faudzi, Y. (2010). Prevalence of diabetes in the Malaysian National Health Morbidity Survey III 2006. *Medical Journal of Malaysia, 65*(3), 173-179.

Lim, S. Y., Ham, C. R., Park, S. Y., Kim, S., Park, M. R., Jeon, K.,..., Suh, G. Y. (2011). Validation of the Simplified Acute Physiology Score 3 Scoring System in a Korean Intensive Care Unit. *Yonsei Medical Journal, 52*(1), 59-64.

Lin, C. Y., Tsai, F. C., Tian, Y. C., Jenq, C. C., Chen, Y. C., & Fang, J. T. (2007). Evaluation of outcome scoring systems for patients on extracorporeal membrane oxygenation. *The Annals of Thoracic Surgery, 84*(4), 1256-1262.

Lipshutz, A. K. M., Feiner, J. R., Grimes, B., & Gropper, M. A. (2016). Predicting mortality in the intensive care unit: a comparison of the University Health Consortium expected probability of mortality and the Mortality Prediction Model III. *Journal of Intensive Care, 4*(1), 35. doi: 10.1186/s40560-016-0158-z.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14-23. doi: 10.1002/widm.8

Long, J. S. (1997). *Advanced quantitative techniques in the social sciences series, Vol. 7. Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications.

Long, W. J., Griffith, J. L., Selker, H. P., & D'Agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers in Biomedical Research, 26*(1), 74-97. doi: S0010480983710050 [pii].

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337. doi: 10.1023/a:1008929526011.

Ma, Q.-B., Fu, Y.-W., Feng, L., Zhai, Q.-R., Liang, Y., Wu, M., & Zheng, Y.-A. (2017). Performance of Simplified Acute Physiology Score 3 in Predicting Hospital Mortality in Emergency Intensive Care Unit. *Chinese Medical Journal, 130*(13), 1544-1551. doi: 10.4103/0366-6999.208250.

Markgraf, R., Deutschinoff, G., Pientka, L., & Scholten, T. (2000). Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. *Critical Care Medicine, 28*(1), 26-33.

Markgraf, R., Deutschinoff, G., Pientka, L., Scholten, T., & Lorenz, C. (2001). Performance of the score systems Acute Physiology and Chronic Health Evaluation II and III at an interdisciplinary intensive care unit, after customization. *Critical Care, 5*(1), 31-36.

Martin-Sanchez, J. C., Cleries, R., Lidon, C., Gonzalez-de Paz, L., Lunet, N., & Martinez-Sanchez, J. M. (2016). Bayesian prediction of lung and breast cancer mortality among women in Spain (2014-2020). *Cancer Epidemiology, 43*, 22-29. doi: 10.1016/j.canep.2016.05.009.

McClish, D. K., & Powell, S. H. (1989). How well can physicians estimate mortality in a medical intensive care unit? *Medical Decision Making, 9*(2), 125-132.

McGinn, T. G., Guyatt, G. H., Wyer, P. C., Naylor, C. D., Stiell, I. G., & Richardson, W. S. (2000). User's guides to the medical literature: Xxii: how to use articles about clinical decision rules. *Journal of the American Medical Association, 284*(1), 79-84.

Mehta, C. R., & Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine, 14*(19), 2143-2160.

Metnitz, B., Schaden, E., Moreno, R., Le Gall, J. R., Bauer, P., & Metnitz, P. G. (2009). Austrian validation and customization of the SAPS 3 Admission Score. *Intensive Care Medicine, 35*(4), 616-622. doi: 10.1007/s00134-008-1286-2.

Metnitz, P. G., Moreno, R. P., Almeida, E., Jordan, B., Bauer, P., Campos, R. A. (2005). SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive Care Medicine, 31*(10), 1336-1344.

Metnitz, P. G., Valentin, A., Vesely, H., Alberti, C., Lang, T., Lenz, K. (1999). Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. *Intensive Care Medicine, 25*(2), 192-197.

Metnitz, P. G. H., Lang, T., Vesely, H., Valentin, A., & Le Gall, J. R. (2000). Ratios of observed to expected mortality are affected by differences in case mix and quality of care. *Intensive Care Medicine, 26*(10), 1466-1472. doi: 10.1007/s001340000638.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machine. *The Journal of Chemical Physics, 21*, 1087-1091.

Meyfroidt, G., Güiza, F., Ramon, J., & Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. *Best practice & research. Clinical anaesthesiology, 23*(1), 127-143. doi: 10.1016/j.bpa.2008.09.003.

Mickey, R. M., & Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology, 129*(1), 125-137.

Mohammadzadeh, F., Noorkojuri, H., Pourhoseingholi, M. A., Saadat, S., & Baghestani, A. R. (2014). Predicting the probability of mortality of gastric cancer patients using decision tree. *Irish Journal of Medical Science*. doi: 10.1007/s11845-014-1100-9.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine, 162*(1), W1-73. doi: 10.7326/M14-0698.

Moralez, G. M., Rabello, L. S. C. F., Lisboa, T. C., Lima, M. d. F. A., Hatum, R. M., De Marco, F.V.C.,...,ORCHESTRA Study Investigators. (2017). External validation of SAPS 3 and MPM(0)-III scores in 48,816 patients from 72 Brazilian ICUs. *Annals of Intensive Care, 7*, 53. doi: 10.1186/s13613-017-0276-3.

Moreno, R. P., & Apolone, G. (1997). Impact of different customization strategies in the performance of a general severity score. *Critical Care Medicine, 25*(12), 2001-2008.

Moreno, R. P., & Matos, R. (2000). The "new" scores: what problems have been fixed, and what remain? *Current Opinion in Critical Care, 6*(3), 158-165.

Moreno, R. P., Metnitz, P. G., Almeida, E., Jordan, B., Bauer, P., Campos, R. A. (2005). SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Medicine, 31*(10), 1345-1355.

Moreno, R. P., Metnitz, P. G., Metnitz, B., Bauer, P., Afonso de Carvalho, S., & Hoechtl, A. (2008). Modeling in-hospital patient survival during the first 28 days after intensive care unit admission: a prognostic model for clinical trials in general critically ill patients. *Journal of Critical Care, 23*(3), 339-348. doi: S0883-9441(07)00188-8 [pii]10.1016/j.jcrc.2007.11.004.

Moreno, R. P., Miranda, D. R., Fidler, V., & Van Schilfgaarde, R. (1998). Evaluation of two outcome prediction models on an independent database. *Critical Care Medicine, 26*(1), 50-61.

Moreno, R. P., & Morais, P. (1997). Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. *Intensive Care Medicine, 23*(2), 177-186.

Mullett, C. J., Evans, R. S., Christenson, J. C., & Dean, J. M. (2001). Development and impact of a computerized pediatric antiinfective decision support program. *Pediatrics*, 108, E175.

Nahra, R., Schorr, C., & Gerber, D. R. (2005). Pre-Intensive Care Unit Length of Stay and Outcome in Critically Ill Patients *Chest, 128*(4), 298S.

Nassar Junior, A. P., Mocelin, A. O., Andrade, F. M., Brauer, L., Giannini, F. P., Nunes, A. L. (2013). SAPS 3, APACHE IV or GRACE: which score to choose for acute coronary syndrome patients in intensive care units? *Sao Paulo Medical Journal, 131*(3), 173-178. doi: S1516-31802013000300173 [pii].

Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology, 9*, 56-56. doi: 10.1186/1471-2288-9-56.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*: Academic Press.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New Jersey: Wiley.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology, 229*(1), 3-8.

Pappachan, J. V., Millar, B., Bennett, E. D., & Smith, G. B. (1999). Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. *Chest, 115*(3), 802-810.

Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing, 43*(2), 154-164. doi: 10.4040/jkan.2013.43.2.154.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379. doi: S0895-4356(96)00236-3 [pii].

Pendergraft, T. B., Stanford, R. H., Beasley, R., Stempel, D. A., & McLaughlin, T. (2005). Seasonal variation in asthma-related hospital and intensive care unit admissions. *J Asthma, 42*(4), 265-271.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research, 96*(1), 3-14. doi: 10.1080/00220670209598786.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: a learning-curve analysis. *The Journal of Machine Learning Research, 4*, 211-255.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of Medical Systems, 26*(5), 445-463.

Purcell, G. P. (2005). What makes a good clinical decision support system. *British Medical Journal, 330*, 740 - 741.

R Development Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Raftery, A., & Lewis, S. (1992). How many iterations in the Gibbs sampler. In J. M. Bernardo, J. Berger, A. P. Dawid & J. F. M. Smith (Eds.), *Bayesian Statistics 4* (Vol. 4, pp. 763-774): Claredon Press, Oxford.

Randolph, A. G., Guyatt, G. H., Calvin, J. E., Doig, G., & Richardson, W. S. (1998). Understanding articles describing clinical prediction tools. Evidence Based Medicine in Critical Care Group. *Critical Care Medicine, 26*(9), 1603-1612.

Riviello, E. D., Kiviri, W., Fowler, R. A., Mueller, A., Novack, V., Banner-Goodspeed, V. M.,...,Twagirumugabe, T. (2016). Predicting Mortality in Low-Income Country ICUs: The Rwanda Mortality Probability Model (R-MPM). *PLoS ONE, 11*(5), e0155858.

Rokach, L., & Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*: World Scientific Publishing Co., Inc.

Rothen, H. U., Stricker, K., Einfalt, J., Bauer, P., Metnitz, P. G., Moreno, R. P. (2007). Variability in outcome and resource use in intensive care units. *Intensive Care Medicine, 33*(8), 1329-1336. doi: 10.1007/s00134-007-0690-3.

Rowan, K. M., Kerr, J. H., Major, E., McPherson, K., Short, A., & Vessey, M. P. (1994). Intensive Care Society's Acute Physiology and Chronic Health Evaluation (APACHE II) study in Britain and Ireland: a prospective, multicenter, cohort study comparing two methods for predicting outcome for adult intensive care patients. *Critical Care Medicine, 22*(9), 1392-1401.

Rudolfer, S. M., Paliouras, G., & Peers, I. S. (1999). A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. *Computers and Biomedical Research, 32*(5), 391-414. doi: 10.1006/cbmr.1999.1521 [doi] S0010480999915215 [pii].

Rue, M., Quintana, S., Alvarez, M., & Artigas, A. (2001). Daily assessment of severity of illness and mortality prediction for individual patients. *Critical Care Medicine, 29*(1), 45-50.

Ryynänen, O.-P., Soini, E. J., Lindqvist, A., Kilpeläinen, M., & Laitinen, T. (2013). Bayesian predictors of very poor health related quality of life and mortality in patients with COPD. *BMC Medical Informatics and Decision Making, 13,* 34. doi: 10.1186/1472-6947-13-34.

Schwartz, S., & Cullen, D. J. (1981). How many intensive care beds does your hospital need? *Critical Care Medicine, 9*(9), 625-629.

Scott, I. A., Brand, C. A., Phelps, G. E., Barker, A. L., & Cameron, P. A. (2011). Using hospital standardised mortality ratios to assess quality of care--proceed with extreme caution. *Medical Journal of Australia, 194*(12), 645-648. doi: sco10527_fm [pii].

Siegelaar, S. E., Devries, J. H., & Hoekstra, J. B. (2010). Patients with diabetes in the intensive care unit; not served by treatment, yet protected? *Critical Care, 14*(2), 126-126. doi: 10.1186/cc8881.

Siegelaar, S. E., Hickmann, M., Hoekstra, J. B. L., Holleman, F., & DeVries, J. H. (2011). The effect of diabetes on mortality in critically ill patients: a systematic review and meta-analysis. *Critical Care, 15*(5), R205-R205. doi: 10.1186/cc10440.

Silverstein, M. D. (1988). Prediction instruments and clinical judgment in critical care. *Journal of the American Medical Association, 260*(12), 1758-1759.

Sirio, C. A., Tajimi, K., Tase, C., Knaus, W. A., Wagner, D. P., Hirasawa, H. (1992). An initial comparison of intensive care in Japan and the United States. *Critical Care Medicine, 20*(9), 1207-1215.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 55*(1), 3-23.

Soares, M., & Salluh, J. I. (2006). Validation of the SAPS 3 admission prognostic model in patients with cancer in need of intensive care. *Intensive Care Medicine, 32*(11), 1839-1844. doi: 10.1007/s00134-006-0374-4.

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130-135. doi: 10.11919/j.issn.1002-0829.215044.

Souza, A. D. P., & Migon, H. S. (2004). Bayesian binary regression model: an application to in-hospital death after AMI prediction. *Pesquisa Operacional, 24*, 253-267.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639. doi: 10.1111/1467-9868.00353.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.), 21*(1), 128-138. doi: 10.1097/EDE.0b013e3181c30fb2.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine, 18*(10), 1099-1104. doi: 10.1111/j.1553-2712.2011.01185.x

Stone, C. J., & Koo, C. Y. (1985). *Additive splines in Statistics.* Paper presented at the Proceedings of the Statistical Computing Section, Washington, D.C.

Teres, D., & Lemeshow, S. (1999). When to customize a severity model. *Intensive Care Medicine, 25*(2), 140-142.

Toft, N., Innocent, G. T., Gettinby, G., & Reid, S. W. J. (2007). Assessing the convergence of Markov Chain Monte Carlo methods: An example from evaluation of diagnostic tests in absence of a gold standard. *Preventive Veterinary Medicine, 79*(2), 244-256. doi: 10.1016/j.prevetmed.2007.01.003.

Tong, J. M. G., Tai, L. L., Tan, C. C., Ahmad, S., Asniza, & Lim, C. H. (2012). *Malaysian Registry of Intensive Care 2012 Report.* Retrieved from www.mric.org.my.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian Estimation of Item Response Curves. *Psychometrika, 51*, 251-267.

van Gestel, Y. R., Lemmens, V. E., Lingsma, H. F., de Hingh, I. H., Rutten, H. J., & Coebergh, J. W. (2012). The hospital standardized mortality ratio fallacy: a narrative review. *Medical Care, 50*(8), 662-667. doi: 610.1097/MLR.1090b1013e31824ebd31829f.

Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C., & Habbema, J. D. F. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology, 58*(5), 475-483.

Vincent, J.-L., Preiser, J.-C., Sprung, C. L., Moreno, R., & Sakr, Y. (2010). Insulin-treated diabetes is not associated with increased mortality in critically ill patients. *Critical Care, 14*(1), R12-R12. doi: 10.1186/cc8866.

Vincent, J. L., de Mendonca, A., Cantraine, F., Moreno, R., Takala, J., Suter, P. M. (1998). Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on "sepsis-related problems" of the European Society of Intensive Care Medicine. *Critical Care Medicine, 26*(11), 1793-1800.

Wagner, D. P., Knaus, W. A., & Draper, E. A. (1983). Statistical validation of a severity of illness measure. *American Journal of Public Health, 73*(8), 878-884.

Wagner, D. P., Knaus, W. A., Harrell, F. E., Zimmerman, J. E., & Watts, C. (1994). Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. *Critical Care Medicine, 22*(9), 1359-1372.

Wang, W., Song, X.-T., Chen, Y.-D., Yang, X.-S., Xu, F., Zhang, M. (2016). Prediction of risk of cardiovascular events in patients with mild to moderate coronary artery lesions using naïve Bayesian networks. *Journal of Geriatric Cardiology : JGC, 13*(11), 899-905. doi: 10.11909/j.issn.1671-5411.2016.11.004.

Warner, R. (2012). *Applied Statistics: From Bivariate through Multivariate Techniques*. Thousand Oaks: Sage Publications, Inc.

Wong, D. T., Crofts, S. L., Gomez, M., McGuire, G. P., & Byrick, R. J. (1995). Evaluation of predictive ability of APACHE II system and hospital outcome in Canadian intensive care unit patients. *Critical Care Medicine, 23*(7), 1177-1183.

Wong, D. T., & Knaus, W. A. (1991). Predicting outcome in critical care: the current status of the APACHE prognostic scoring system. *Canadian Journal of Anaesthesia, 38*(3), 374-383. doi: 10.1007/BF03007629.

Wong, H. R., Lindsell, C. J., Pettila, V., Meyer, N. J., Thair, S. A., Karlsson, S. (2014). A multibiomarker-based outcome risk stratification model for adult septic shock. *Critical Care Medicine, 42*(4), 781-789. doi: 10.1097/CCM.0000000000000106.

Xing, X., Gao, Y., Wang, H., Huang, C., Qu, S., Zhang, H.,..., Sun, K.L. (2015). Performance of Three Prognostic Models in Patients with Cancer in Need of Intensive Care in a Medical Center in China. *PLoS ONE, 10*(6), e0131329.

Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *9*, 79-94. doi: 10.20982/tqmp.09.2.p079.

Zhang, J., Goode, K. M., Rigby, A., Balk, A. H., & Cleland, J. G. (2013). Identifying patients at risk of death or hospitalisation due to worsening heart failure using decision tree analysis: evidence from the Trans-European Network-Home-Care Management System (TEN-HMS) study. *International Journal of Cardiology, 163*(2), 149-156. doi: S0167-5273(11)00534-1[pii]10.1016/j.ijcard.2011.06.009.

Zimmerman, J. E., Alzola, C., & Von Rueden, K. T. (2003). The use of benchmarking to identify top performing critical care units: a preliminary assessment of their policies and practices. *Journal of Critical Care, 18*(2), 76-86. doi: 10.1053/jcrc.2003.50005.

Zimmerman, J. E., Knaus, W. A., Judson, J. A., Havill, J. H., Trubuhovich, R. V., Draper, E. A. (1988). Patient selection for intensive care: a comparison of New Zealand and United States hospitals. *Critical Care Medicine, 16*(4), 318-326.

Zimmerman, J. E., & Kramer, A. A. (2008). Outcome prediction in critical care: the Acute Physiology and Chronic Health Evaluation models. *Current Opinion in Critical Care, 14*(5), 491-497.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., & Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Critical Care Medicine, 34*(5), 1297-1310.

Zimmerman, J. E., Wagner, D. P., Draper, E. A., Wright, L., Alzola, C., & Knaus, W. A. (1998). Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. *Critical Care Medicine, 26*(8), 1317-1326.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Partial findings in this thesis were published in the following Thomson Reuters ISI-indexed journals:-

1. Wong, R.S.Y., Ismail, N.A. (2016). An Application of Bayesian Approach in Modeling Risk of Death in an Intensive Care Unit. *PLoS ONE*, *11*(3), e0151949. doi:10.1371/journal.pone.0151949.

2. Wong, R.S.Y., Ismail, N.A., Tan, C.C. (2015). An External Independent Validation of APACHE IV in a Malaysian Intensive Care Unit. *Annals Academy of Medicine Singapore*, *44*(4), 127-132.